A LEADER-FOLLOWER PARTIALLY OBSERVED MARKOV GAME

A Thesis
Presented to
The Academic Faculty

by

Yanling Chang

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Industrial and Systems Engineering

Georgia Institute of Technology
December 2015

A LEADER-FOLLOWER PARTIALLY OBSERVED MARKOV GAME

Approved by:

Professor Chelsea C. White III,
Co-Advisor
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Alan L. Erera, Co-Advisor
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Turgay Ayer
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Enlu Zhou
School of Industrial and Systems
Engineering
Georgia Institute of Technology

Professor Luca Dieci
School of Mathematics
Georgia Institute of Technology

Date Approved: October 22, 2015

To my family and true friends

for their unwavering support.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my two advisors, Professor Chelsea C. White III and Professor Alan L. Erera, for their constant exceptional guidance, support and encouragement. I had no support and no prospects when I first enrolled as a PhD student in ISyE. Professor White and Professor Erera had the heart to take me in as a graduate research assistant, and gave me the opportunity to work with them on fantastic research problems. Throughout the years of my Ph.D. study, they always provided me their constructive comments and insights to help me better understand how to approach the research problems, and guided me to learn to think more critically and creatively as an independent researcher. They also have been great examples in the way they treat with others and maintain their personal lives. I would like to extend my special gratitude to Professor White who spent so many hours on my papers and proofs, kept inspiring me new research ideas, and showed me how to work with others and how to become an excellent instructor and researcher. He is truly instrumental in guiding my research. It has been a great honor and pleasure to have two such great advisors in my Ph.D. study.

I will forever be thankful to Professor Dieci who laid down a solid mathematical background for me. I sincerely appreciate the beauty of mathematics he presented to me which has inspired me on some new research ideas. I would also like to express my gratitude to Professor Turgay Ayer and Professor Enlu Zhou for their willingness to serve on my committee without hesitation and their insightful questions and discussions, all of which have contributed to my body of knowledge and to this dissertation. I would like to acknowledge Professor Lauren Berrings Davis who also provided guidance on this dissertation and have served as a role model when we worked together.

Throughout the years spent at Georgia Tech, I appreciate all the supports from my fellow colleagues. I would like to thank Brian Kues and Yu Zhang for their invaluable supports on our projects, critical to the completion of this dissertation. I also would like to thank Niao He, Ran Li, Satya Sarvani Malladi and Fangfang Xiao for their helpful information, discussions and making my life fun. I will never forget the long conversations we have had on anything from research plans and opportunities, seminars, mathematical proofs, history, literature to job hunting. I look forward to continuing these invaluable friendships with these great people.

Last but not least, I would like to thank my family and my husband Qiao for their love, sacrifices and support. They helped instill belief in my talent and abilities when I was frustrated and celebrated my successes, although they don't have much idea on what I was studying. It would have been impossible to finish this dissertation without their love and support.

I also acknowledge National Center For Food Protection and Defense, a Homeland Security Center of Excellence for providing support for this work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

The intent of this dissertation is to generate a set of non-dominated finite-memory policies from which one of two agents (the leader) can select a most preferred policy to control a dynamic system that is also affected by the control decisions of the other agent (the follower). The problem is described by an infinite horizon total discounted reward, partially observed Markov game (POMG). Each agents policy assumes that the agent knows its current and recent state values, its recent actions, and the current and recent possibly inaccurate observations of the other agents state. For each candidate finite-memory leader policy, we assume the follower, fully aware of the leader policy, determines a policy that optimizes the followers criterion. The leader-follower assumption allows the POMG to be transformed into a specially structured, partially observed Markov decision process that we use to determine the followers best response policy for a given leader policy. We then present a value determination procedure to evaluate the performance of the leader for a given leader policy, based on which non-dominated set of leader polices can be selected by existing heuristic approaches.

We then analyze how the value of the leaders criterion changes due to changes in the leaders quality of observation of the follower. We give conditions that insure improved observation quality will improve the leaders value function, assuming that changes in the observation quality do not cause the follower to change its policy. We show that discontinuities in the value of the leader criterion, as a function of observation quality, can occur when the change of observation quality is significant enough for the follower to change its policy. We present conditions that determine

when a discontinuity may occur and conditions that guarantee a discontinuity will not degrade the leaders performance. This framework has been used to develop a dynamic risk analysis approach for U.S. food supply chains and to compare and create supply chain designs and sequential control strategies for risk mitigation.

CHAPTER I

INTRODUCTION

Models of sequential decision-making under uncertainty provide a rich normative framework for one or more intelligent decision-makers to improve, e.g., optimize, the operation of a system subject to control over a horizon containing a sequence of decision epochs. The solutions of such models can provide guidance as to how decision-makers should select actions, based on currently available data, in order to achieve their objectives. At each decision epoch, each decision-maker selects an action that: (1) accrues an immediate reward dependent on the current state of the system and the current actions taken by all of the decision-makers and (2) affects what the state of the system will be at the next decision epoch. Each decision-maker must balance how its decision has impact on the immediate reward accrued and how its decision will affect rewards to be accrued in the future, based on all currently available data. If there are multiple decision-makers, each decision-maker must also consider how the decisions of the other decision makers will have impact on its ability to achieve its goals.

State observations on which decision-makers base their decisions may be inaccurate or imperfect measures of the system state. There are two extremes of observation quality:

- The completely observed case, where the observation data provide an accurate description of the current state of the system;
- The completely unobserved case, where the observation data are unrelated to the current state of the system and hence are uninformative.

Models of sequential decision-making that assume the state is completely observed are typically considerably more computationally tractable than models of sequential decision-making that assume the state is partially observed. However, partially observed problem formulations often represent more realistic models of real world decision-making situations.

With respect to the number of decision-making agents and the accuracy of state observations available to the agents, there are four possible cases:

- Single-agent completely observed decision-making;
- Multi-agent completely observed decision-making;
- Single-agent partially observed decision-making;
- Multi-agent partially observed decision-making.

Each of these four cases will be described and its literature will be reviewed in the following subsections. This section will then conclude with the outline and main contributions of the dissertation.

1.1 Completely Observed Single-Agent Decision Making

The Markov decision process (MDP) is an important and well-studied optimization model of discrete stage, completely observed single agent sequential decision making in a stochastic environment. The MDP has been widely applied to problems in automated control, manufacturing processes, healthcare, etc (Banerjee and Gupta 2013; Li and Sun 2013; Peters et al. 2015; Moulik et al. 2014). Many rich and elegant theoretical and computational advances have been achieved over the past decades.

A MDP consists of decision epochs, a state space, an action space, a reward structure, a criterion, and transition probabilities. A formal definition of a MDP can be

found in many standard textbooks (Bellman, 1957; Bertsekas, 1987; Denardo, 1982; Howard 1960; Puterman 1994). The expected total discounted reward criterion for the infinite horizon is the focus throughout this dissertation. Details on average expected reward models and recent developments can be found in White and Scherer (1987), Puterman (1994), Cavazos-Cadena (1989, 1992), Feinberg and Park (1994), Meyn (1997), Bertsekas (1998), Lewis and Puterman (2001), and Shlakhter (2010).

Standard solution procedures for the infinite horizon discounted case include value iteration and its variants, (modified) policy iteration (with action elimination), and linear programming (Puterman 1994). Reward revision (White, Thomas and Scherer 1985) is another technique that can solve the infinite horizon problem efficiently by constructing a sequence of MDPs that share the same fixed point as the original problem. The sequence of MDPs is constructed by periodic revisions of its reward structure.

Recent algorithms for large scale MDPs aim to improve the performance of value and policy iteration by eliminating redundant or useless backup steps and/or generating a good backup ordering so that the transition matrix can be in a computationally useful form (for example, triangular matrices) (Barto, Bradtke and Singh, 1995; Hansen and Zilberstein 2001; Wingate and Seppi, 2005; McMahan and Gordon, 2005; Sanner et al., 2009; Dai, Weld and Goldsmith 2011). Another type of solution technique is based on state aggregation to approximate the original problem by a MDP having a smaller state space and/or action space (White and White 1989; Roy 2006; Jia 2011).

The multiobjective MDP is an extension of the MDP where there are multiple, possibly conflicting objectives under consideration. Chatterjee, Majumdar and Henzinger (2006) showed that the Pareto optimal can be achieved by a randomized memoryless

policy and the Pareto frontier can be approximated in polynomial time. Viswanathan, Aggarwal and Nair (1977), Hernandez-Lerma and Romera (2004) and Chatterjee (2006) reformulate the multiobejctive MDP as a multi-objective linear programming. White and Kim (1980) solved the problem by reformulating it as a specially structured partially observed MDP. Wakuta (1995) combined Fourier elimination with a policy iteration algorithm to compute all optimal stationary policies. Futher discussion of the multiobjective MDP can be found in Roijers et al. (2013).

For many applications, the rewards and the transition probabilities can be very difficult to quantify precisely (Wiesemann, Kuhn and Rustem 2013). For parameter imprecision associated with the reward structure, White and EL-Deib (1986) studied MDPs having a reward function that is affine in an imprecise parameter. Both a successive approximations procedure for the finite horizon case and a policy iteration procedure for the infinite horizon problem were developed. Tan and Hartman (2011) studied how multiple parameters in the reward function can vary while maintaining the optimality of the current solution.

For parameter imprecision associated with imprecise transition probabilities, White and EL-Deib (1994) described the imprecise transition probabilities by a finite number of linear inequalities and developed solution procedures based on successive approximations, reward revision, and modified policy iteration. Nilim and Ghaoui (2005) characterized imprecise transition structures in terms of nonconvex sets, presented a duality result, and demonstrated that the problem can be solved by robust dynamic programming. Delage and Mannor (2010) presented a chance-constrained formulation of the MDP with imprecise parameters in either the reward structure or transition probabilities and studied the effect of this uncertainty on the criterion value. Other related work can be found in Iyengar (2005) and Delgado, Sanner and De Barros (2011).

1.2 Completely Observed Multi-Agent Decision Making

Markov games, also called stochastic games, are generalizations of MDPs that involve a group of agents (also called players) such as a team of robots or several competitors (or collaborators) with different objectives. The Markov game was first introduced by Shapley (1953). At each decision epoch, each player selects an action simultaneously with and independently from all other players. The game evolves to a new state depending on the current state and the joint actions of all decision makers. The Markov game model extends the MDP and game-theoretic frameworks to determine policies for players in which rewards and transitions are determined by the simultaneous actions of all players. This type of game requires no hidden information and all players have complete and perfect information all the time.

The overwhelming focus for the Markov game is on the Markov perfect equilibrium (MPE), an equilibrium where players' policies are Markov policies. Rogers (1969) and Sobel (1971) established an existence result for the Markov game with a finite number of states and actions. For this type of game, they showed that the Markov game has an equilibrium in stationary Markov policies via a routine application of Kakutani's theorem. This existence theorem has been extended to countable state spaces in Parthasarathy (1982) and Rieder (1979). Chakrabarti (1999) proved the existence of the MPE and the semi-MPE for a Markov game having a complete separable state space and compact action spaces for each player. Dutta and Sundaram (1992) studied the existence of the MPE for pure policies for the discounted game and the limiting behavior of the MPE as the discount factor tends to unity. The equilibrium existence problem for the general Markov game was reviewed by Dutta and Sundaram (1998). Related properties of MPE can be found in Doraszelski and

Escobar (2010) and Haller and Lagunoff (2000).

A zero-sum Markov game with a finite number of states and actions can be solved by value iteration or policy iteration, analogous to the MDP (Littman 1996). The general-sum Markov game can be solved via nonlinear programming (Filar and Vrieze 1997).

1.3 Partially Observed Single-Agent Decision Making

A partially observable Markov decision process (POMDP) extends the framework of the standard Markov decision process to situations where an agent only has noisy observations of the system state. The POMDP formalism has successfully extended the applications of MDPs to many realistic problems such as machine maintenance, behavioral ecology, network troubleshooting, hostile target identification, and medical diagnosis (Cassandra 1998).

A POMDP is comprised of the state space S, action space A, observation space Z, transition probability matrices $\{P^a\}_{a\in A}$, observation probability matrices $\{Q^a\}$ and reward function $r : S \times A \to \mathbb{R}$ (White 1991). This thesis only considers the case where the state space S, the observation space Z, and the action space A are finite.

The generality of the POMDP has computational implications. Papadimitriou and Tsitsiklis (1987) prove that finding optimal policies for finite-horizon POMDPs is PSPACE-complete, and Madani, Hanks and Condon (1999) show that the existence of optimal policies for infinite-horizon POMDPs is undecidable. Existing classical exact algorithms are developed based on the fact that the value function v_t is piecewise linear and concave with respect to its sufficient statistic (or belief point $x = P(s(t)|I(t)), s(t)$

is the current state and $I(t)$ includes all past observations $z(t), ...z(1)$ and actions $a(t-1), ..., a(0))$ at each decision epoch t (Sondik 1971, 1978).

There are two types of classical exact algorithms. The first type is to systematically construct the value function v_t piece by piece from Γ_{t+1} by successive approximation. This type first identifies a proper subset $\bar{\Gamma}_t$ of Γ_t to approximate the value function v_t and then searches for the belief points x where this approximation is not accurate. These belief points are often called witness points W. The value function v_t is updated at these witness points, and the resulting optimal γ vectors are added to $\bar{\Gamma}_t$. Thus, the cardinality of $\bar{\Gamma}_t$ is strictly increasing until $\bar{\Gamma}_t = \Gamma_t$. Since there are only finitely many γ vectors in Γ_t, the algorithm stops after a finite number of iterations (Cassandra 1994). Algorithms of this type include Sondik's one-pass algorithm (Sondik 1971, Smallwood and Sondik 1973), Cheng's algorithms (Cheng 1988), and the witness algorithm (Littman 1994).

The second type of approach first constructs all possible γ vectors that could describe the value function in order to form a superset $\bar{\Gamma}_t, \bar{\Gamma}_t \supseteq \Gamma_t$ (Cassandra 1994). A vector γ that is necessary to define the value function is called a defining vector. The smallest set of defining vectors Γ_t is often called the minimal representation of the value function (Lin, Bean and White, 1998, 2004). The main objective of this type of algorithm is to effectively identify these defining vectors and remove all redundant vectors using a so-called **PURGE** operator (Lin, Bean and White 1998, 2004). A geometric interpretation of the **PURGE** operator is to identify extreme points of a convex hull of a point set to find the minimum representation of the value function (Zhang 2010). Algorithms of this type include the Monahan/Lark algorithm (Monahan 1982, White 1991), the (Generalized) Incremental pruning algorithm (Cassandra, Littman, Zhang 1997; Naser-Moghadasi 2010, 2012) and the Hybrid genetic/optimization algorithm

(Lin, Bean and White 1998, 2004). The latter two algorithms are used in this thesis to solve POMDP problems.

A finite state controller is comprised of a finite set of internal states (called "control states"), action rules and transition rules (Zhang 2010). Hansen's policy iteration algorithm is designed to search for an optimal finite state controller (Hansen 1998). This algorithm is guaranteed to converge to an ϵ-optimal finite-state controller in a finite number of iterations, and it will be optimal if the finite state controller cannot be further improved.

A finite memory controller is a suboptimal design. It is a mapping from the set of finite recent histories $h(t,\tau)$ into the action space A, where τ is a design parameter for the maximum memory length and $h(t,\tau) = \{z(t), a(t-1), z(t-1), a(t-2), ..., z(t-\tau+1), a(t-\tau)\}$. The corresponding γ vector for a finite memory controller has been computed by Ortiz, Erera and White (2013) and White and Scherer(1994).

A finite memory controller can be easily represented as a finite state controller in which an internal state corresponds to a finite recent history $h(t-\tau)$. A finite memory controller can be in some cases an optimal finite state controller. White and Scherer (1994) constructed lower and upper bounds for value functions of finite memory controllers and showed that a finite memory controller is an optimal finite state controller if $P(h(t,\tau))$ is of rank 1. However, there are examples that show a finite state controller cannot be represented by any finite memory controller $\forall\tau < \infty$ (Yu 2007).

The point-based POMDP, first suggested by Lovejoy (1991), is considered a very important contribution to POMDP research because it can approximately solve large

POMDPs rapidly. It maintains an approximate value function over a finite subset of the belief space in order to avoid the exponential growth of γ vectors that define a value function. An extensive literature review of point-based POMDP algorithms can be found in Shani, Pineau and Kaplow (2013). Many point-based algorithms can continually improve the approximation of the value function over time and can be terminated to obtain an approximate near optimal policy based on a given time constraint.

1.4 Partially Observed Multi-agent Decision Making

Partially observed multi-agent decision making frameworks involve multiple intelligent and adaptive decision makers, each of which can choose actions that affect the dynamics of the system, based on current and past possibly inaccurate state observations.

The I-POMDP is a generalization of the POMDP to a multiagent setting. The I-POMDP defines the possible models of all agents and includes these agents in the description of the system (Gmytrasiewicz and Doshi 2005). The I-POMDP extends the concept of the system state to include the other agents' models, and these extended states are called interactive states (Gmytrasiewicz and Doshi 2005). In a manner similar to the POMDP, the agent's belief over these interactive states is a sufficient statistic and the value function is piecewise linear and convex with respect to this belief. However, a belief over interactive states is also a part of other agents' models. Hence, the belief is infinitely nested so that the I-POMDP can not be solved exactly (Gmytrasiewicz and Doshi 2005). Current solution techniques for the I-POMDP include policy iteration (Sonu and Doshi 2012), point based value iteration (Doshi and Perez, 2008), and interactive particle filtering (Doshi and Gmytraslewicz 2009).

The decentralized partially observable Markov decision process (DEC-POMDP) is a framework for planning for groups of cooperative agents in a stochastic and partially observable environment, where all the agents share the same reward function. Contrary to the single agent POMDP, the DEC-POMDP in general cannot provide the entire information needed for the whole group to select an action, although the agents are collaborating with each other. The reason is that an agent only knows its local observation Z^i but not the complete observation vector $Z = \{Z^i, 1 \leq i \leq N\}$($N$ is the number of agents), and each agent has to choose its action a^i based on its own history of actions taken and observations received so far. The goal of the DEC-POMDP is to search for a joint policy $\pi = \{\pi^1, ...\pi^N\}$ which maximizes the total discounted reward for the group. However, solving finite horizon DEC-POMDPs is provably NEXP-complete (Bernstein et al. 2002) and even computing solutions with absolutely bounded error is also NEXP-complete (Rabinovich, Goldman and Rosenschein 2003).

One way to solve the DEC-POMDP is to introduce a communication channel to the group of agents so that the agents can communicate with each other about observations and actions. If the agents can communicate without limit and at no cost, then every agent can receive observations and actions from all of its teammates at each decision epoch. Consequently, free communication reduces the DEC-POMDP to a centralized single agent decision making problem, which can be solved by standard POMDP techniques. However, in reality communication is not free or unlimited, and a model that assumes free communication might not be realistic. To overcome this issue, Emery-Montemerlo et al. (2004), Roth, Simmons and Veloso (2005) and Nair et al. (2004) first solved the large centralized POMDP as if communication were truly free, and each agent executes the resulting joint policy in a distributed system

with communication constraints; Xuan, Lesser and Zilberstein (2001), Goldman and Zilberstein (2003) and Spaan, Gordon and Vlassis (2006) incorporate the communication decisions into the agent's policy.

The past ten years have enjoyed significant success in the development of both exact algorithms and approximation algorithms for the DEC-POMDP. For the finite horizon DEC-POMDP, existing exact algorithms include dynamic programming (Hansen, Bernstein and Zilberstein 2004; Boularias and Chaib-Draa 2008; Amato, Dibangoye and Zilberstein 2009), linear programming (Aras and Dutech 2010), and heuristic search (Dibangoye et al. 2013). However, these approaches become quickly intractable due to their complexity. Specifically, the number of policies grows doubly exponentially as the number of decision epochs increases, and hence the models become computationally intractable quickly as the problem size grows. No exact solution has been found yet for the infinite-horizon DEC-POMDP in the current literature (Bernstein et al. 2009; Dibangoye, Mouaddib and Chaib-Draa 2011). A detailed review of existing DEC-POMDP models, solution procedures and complexity results can be found in Seuken and Zilberstein (2005), Oliehoek (2012) and Amato et al. (2013).

The DEC-POMDP is a cooperative version of a partially observed stochastic game (POSG) where each agent may have different objectives. The POSG framework is very general and very challenging computationally. So far, the POSG is still a relatively unexamined research area, and the literature on POSGs is relatively sparse. Hansen, Bernstein and Zilberstein (2004) have pointed out that it is not possible to solve a POSG by transforming it into a completely observable stochastic game over the belief states.

A dynamic programming approach (Hansen, Bernstein and Zilberstein 2004) was developed to produce optimal policies for cooperative POSGs, i.e., the DEC-POMDPs. However, the size of the policy sets for each agent increases doubly exponentially in the horizon t, $|A_i|^{|O_i|^t}$, so that the dynamic programming approach cannot solve problems with large horizons (Hansen, Bernstein and Zilberstein 2004). A bounded approximation technique has been attempted by Kumar and Zilberstein (2009) to scale the POSG by several orders of magnitude. However, the dynamic programming approach still cannot be used to compute optimal policies for general POSGs.

Because of the difficulty in solving the general POSG, much work has been done on POSGs with special structure. The two-player zero-sum partially observed stochastic game is an active research area. Ghosh, McDonald and Sinha (2004) transformed a zero-sum stochastic game with partial information into an equivalent problem with complete information. Saha (2014) considered a partially observable zero-sum stochastic game with average payoff criterion. A survey of zero-sum POSGs can be found in Chatterjee, Doyen and Henzinger(2013). Dermed, Isbell and Weiss (2011) approximated POSGs by Markov games of incomplete information, which can be converted into a completely observed stochastic game. Approximate algorithms for the infinite horizon general POSG have not been presented in the literature.

1.5 Outline and Contributions

This dissertation presents and analyzes models of a sequential stochastic game involving two decision-makers. These models extend the existing literature on stochastic games by explicitly considering the strategic interaction over time of two non-myopic agents, a leader and a follower, each of whom adjust its decisions according to the other agent's decisions for the case where neither agent has complete information

about the other agent. The agents can be cooperative, non-cooperative, or a mixture of both. Multi-objective optimization is also introduced to enable multi-attribute decision making. This dissertation also contributes to the analysis of the value of information for this class of sequential stochastic games in order to better inform the decision to seek or not seek improved state observation quality and what resulting changes in performance to expect if state observation quality is improved in the multi-agent decision making framework. We apply our results to a situation involving the management of a food processing facility subject to a possible attack from an individual (or group of individuals) intent on contaminating the food with a biological or chemical toxin. Thus, the results presented in this dissertation also contribute to how risk can be measured and managed, assuming two intelligent and adaptive decision-makers, and hence to the risk and decision analysis literature.

The dissertation is organized into three main chapters. The first of these chapters (Chapter II) presents a model of and analyzes a leader-follower, multi-objective partially observed infinite horizon Markov game, where it is assumed that the follower selects its policy with complete knowledge of the policy selected by the leader. We show how the results of this POMG can be used to support decision-making involving a leader having multiple objectives. The second of these chapters (Chapter III) considers the single objective version of the problem considered in Chapter II and investigates the impact of how accurately the leader observes the follower's state on the performance of the leader, thus representing an analysis of the value of information for this class of POMGs. The third of these chapters (Chapter IV) applies the results of the first two chapters in order to quantify the risk of a food production facility to an intelligent and adaptive adversary intent on delivering a chemical or biological toxin to the general population through use of the food supply chain. The goal of this chapter is to develop a new model of dynamic risk analysis that can explicitly

describe the strategic interaction between two intelligent and adaptive agents with different objectives, and to provide decision support to the defender as to when and what action should be taken in order to achieve the defender's (possibly multiple) objectives.

We now provide a more detailed description of the results found in these three chapters. Chapter II presents a model of and analyzes a leader-follower, multi-objective partially observed infinite horizon Markov game, where it is assumed that the follower selects its policy with complete knowledge of the policy selected by the leader. The objective is to generate a set of non-dominated policies from which the leader can select a most preferred policy in order to control a dynamic system that is also affected by the control decisions of the follower. The leader-follower assumption allows the POMG to be converted into a specially structured partially observed Markov decision process (POMDP) for the follower. We present a value determination procedure to evaluate the performance of a given leader policy. This performance evaluation process provides a foundation to generate non-dominated leader's policies by using existing heuristic approaches such as the multi-objective genetic algorithm (MOGA). Treating performance measures as fitness measures, the MOGA creates successive generations of leader policies and eliminates all but the non-dominated set of leader policies. The leader can then select the policy from this set that the leader considers to be the most preferred. This approach to decision-making extends the existing literature on sequential games by explicitly considering the infinite horizon interaction of two non-myopic agents, each of whom adjusts its decisions according to the other agent's decisions. Furthermore, this model considers the case where neither agent has complete knowledge of the other agent's current state.

In Chapter III, we investigate the value of information for the leader for the single objective version of the partially observed Markov game (POMG) developed in Chapter II. We first summarize previously determined results in the literature for the POMDP, based on two different definitions of observation quality and then compare and contrast these two definitions. We determine how the leader's criterion value changes due to changes in the leader's quality of observation of the follower and show that the value of information results for the POMDP can be directly extended to the POMG for sufficiently small changes of observation quality. We show that discontinuities in the leader's value function, as a function of observation quality, can occur when the change of observation quality is significant enough for the follower to change its policy. We also present conditions that determine when a discontinuity may occur and conditions that guarantee that a discontinuity will not degrade leader performance. We show that when the leader and the follower are collaborative and the follower completely observes the leader's initial state, discontinuities in the leader's value function will not occur. However, we present examples that show that improving observation quality does not necessarily improve the leader's value function, whether or not the POMG is a collaborative game.

In Chapter IV, we apply the POMG analyzed in Chapters II and III in order to develop a new dynamic risk analysis model that can explicitly describe the strategic interaction between two intelligent and adaptive agents with different objectives over an at most countable number of decision epochs. Both of the agents can select its action on the basis of possibly inaccurate data collected at current and past decision epochs. Our risk analysis tool is comprised of two components: a consequence assessment tool and a game theoretic optimization model. Several different variations of the model, where these variations are distinguished by the quality of observation that one agent has of the other agent's current state, are considered in order to analyze how the

defender's performance changes as the information accuracy changes. We present our approach by an illustrative example involving a liquid egg production system where an adversary may seek to contaminate the food production facility with a biological or chemical toxin, and the defender (e.g. the plant manager) has to balance achieving two objectives: (1) maximizing plant productivity, and (2) minimizing the expected consequence of deliberate contamination. Our preliminary analysis shows that the system is under greatest risk if the defender's state can be accurately observed by the attacker but the defender can only inaccurately observe the attacker's state. We show the impact on risk reduction of reducing the attacker's observation accuracy of the defender. We evaluate the defender's performance, as a function of the defender's observation accuracy of the attacker, indicating the significant value-added that observation accuracy can play in such situations. We show how system risk can be dynamic as a result of the strategic interaction between two agents. We show that a good defender's policy can redirect the attacker's interests to less vulnerable targets and lengthen the expected time till an attack occurs. A sensitivity analysis is performed to better understand what parameter values need careful assessment and what parameters do not.

1.6 References

[1] Amato, C., Chowdhary, G., Geramifard, A., Ure, N. K., and Kochenderfer, M. J., "Decentralized control of partially observable Markov decision process," in *Proceedings of the Fifty-Second IEEE Conference on Decision and Control*, 2013.

[2] Amato, C., Dibangoye, J. S., and Zilberstein, S., "Incremental policy generation for finite-horizon dec-pomdps," in *Proceedings of the Nineteenth International Conference on Automated Planning and Scheduling*, pp. 2–9, 2009.

[3] Aras, R. and Dutech, A., "An investigation into mathematical programming for finite horizon decentralized POMDPs," *Artificial Intelligence*, vol. 37, pp. 329–396, 2010.

[4] Banerjee, A. G. and Gupta, S. K., "Research in automated planning and control for micromanipulation," *IEEE Transactions on Automation Science and Engineering*, vol. 10, no. 3, pp. 485–495, 2013.

[5] Barto, A. G., Bradtke, S. J., and Singh, S. P., "Learning to act using real-time dynamic programming," *Artificial Intelligence*, vol. 72, pp. 81–138, 1995.

[6] Bellman, R., "A Markovian decision process," *Journal of Mathematics and Mechanics*, vol. 6, no. 5, pp. 679–684, 1957.

[7] Bernstein, D. S., Amato, C., Hansen, E. A., and Zilberstein, S., "Policy iteration for decentralized control of Markov decision processes," *Journal of Artificial Intelligence Research*, vol. 34, pp. 80–132, 2009.

[8] Bernstein, D. S., Givan, R., Immerman, N., and Zilberstein, S., "The complexity of decentralized control of Markov decision processes," *Mathematics of Operations Research*, vol. 27, no. 4, pp. 819–840, 2002.

[9] Bertsekas, D. P., *Dynamic programming*. Englewood Cliffs, NJ: Prentice Hall, 1987.

[10] Bertsekas, D., "A new value iteration method for the average cost dynamic programming problem," *SIAM Journal of Control and Optimization*, vol. 36, pp. 742–759, 1998.

[11] Boularias, A. and Chaib-draa, B., "Exact dynamic programming for decentralized POMDPs with lossless policy compression," in *Proceedings of the*

Eighteenth International Conference on Automated Planning and Scheduling, pp. 20–27, 2008.

[12] Cassandra, A., "Optimal policies for partially observable Markov decision processes," tech. rep., Brown University, Department of Computer Science, Providence RI, 1994.

[13] Cassandra, A., "A survey of POMDP applications," in *Working Notes of AAAI 1998 Fall Symposium on Planning with Partially Observable Markov Decision Processes*, pp. 17–24, 1998.

[14] Cassandra, A., Littman, M., and Zhang, N., "Incremental pruning: a simple, fast, exact method for partially observable Markov decision processes," in *Proceedings Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-97)*, (Morgan Kaufmann, San Francisco, CA), pp. 54 – 61, 1997.

[15] Cavazos-Cadena, R., "Necessary conditions for the optimality equation in average-reward Markov decision processes," vol. 19, pp. 97–112, 1989.

[16] Cavazos-Cadena, R., "Existence of optimal stationary policies in average reward Markov decision processes with a recurrent state," *Journal of Applied Mathematics and Optimization*, vol. 26, pp. 171–194, 1992.

[17] Chakrabarti, S. K., "Markov equilibria in discounted stochastic games," *Journal of Economic Theory*, vol. 85, pp. 294–327, 1999.

[18] Chatterjee, K., Doyen, L., and Henzinger, T. A., "A survey of partial observation stochastic parity games," *Formal Methods in System Design*, vol. 43, no. 2, pp. 268–284, 2013.

[19] Chatterjee, K., Majumdar, R., and Henzinger, T. A., "Markov decision processes with multiple objectives," *Lecture Notes in Computer Science*, vol. 3884, pp. 325–336, 2006.

[20] Cheng, H.-T., *Algorithms for partially observable Markov decision processes*. PhD thesis, University of British Columbia, British Columbia, Canada, 1988.

[21] Dai, P., Weld, D. S., and Goldsmith, J., "Topological value iteration algorithms," *Journal of Artificial Intelligence Research*, vol. 42, pp. 181–209, 2011.

[22] Delage, E. and Mannor, S., "Percentile optimization for Markov decision processes with parameter uncertainty," *Operations Research*, vol. 58, no. 1, pp. 203–213, 2010.

[23] Delgado, K. V., Sanner, S., and de Barros, L. N., "Efficient solutions to factored MDPs with imprecise transition probabilities," *Artificial Intelligence*, vol. 175, pp. 1498–1527, 2011.

[24] Denardo, E. V., *Dynamic programming*. Englewood Cliffs, NJ: Prentice Hall, 1982.

[25] Dermed, L. M., Isbell, C. L., and Weiss, L., "Markov games of incomplete information for multi-agent reinforcement learning," in *Workshops at the 25th AAAI Conference on Artificial Intelligence*, pp. 43–51, 2011.

[26] Dibangoye, J. S., Mouaddib, A.-I., and Chaib-draa, B., "Toward error-bounded algorithms for infinite-hoziron DEC-POMDPs," in *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems*, pp. 947–954, 2011.

[27] Dibangoye, J. S., Amato, C., Buffet, O., and Charpillet, F., "Optimally solving dec-pomdps as continuous-state MDPs: theory and algorithms," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[28] Doraszelski, U. and Escobar, J. F., "A theory of regular Markov perfect equilibria in dynamic stochastic games: Genericity, stability, and purification," *Theoretical Economics*, vol. 5, pp. 369 – 402, 2010.

[29] Doshi, P. and Gmytraslewicz, P., "Monte Carlo sampling methods for approximating interactive pomdps," *Journal of Artificial Intelligence Research*, vol. 34, pp. 297–337, 2009.

[30] Doshi, P. and Perez, D., "Generalized point based value iteration for interactive pomdps," in *Twenty Third Conference on Artificial Intelligence*, vol. 63-68, 2008.

[31] Dutta, P. K. and Sundaram, R., "Markovian equilibrium in a class of stochastic games: existence theorems for discounted and undiscounted models," *Economic Theory*, vol. 2, pp. 197–214, 1992.

[32] Dutta, P. K. and Sundaram, R. K., *The equilibrium existence problem in general Markovian games*, vol. 5 of *Organizations with Incomplete Information*. Cambridge, UK: Cambridge University Press, 1998.

[33] Emery-Montemerlo, R., Gordon, G., Schneider, J., and Thrun, S., "Approximate solutions for partially observable stochastic games with common payoffs," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS)*, pp. 136–143, 2004.

[34] Feinberg, E. A. and Park, H., "Finite state Markov decision models with average reward criteria," *Stochastic Processes and their Applications*, vol. 49, pp. 159–177, 1994.

[35] Filar, J. A. and Vrieze, K., *Competitive Markov decision processes*. New York: Springer, 1997.

[36] Ghosh, M. K., McDonald, D., and Sinha, S., "Zero-sum stochastic games with partial information," *Journal of Optimization Theory and Applications*, vol. 121, no. 1, pp. 99 – 118, 2004.

[37] Gmytrasiewicz, P. and Doshi, P., "A framework for sequential planning in multiagent settings," *Journal of Artificial Intelligence Research*, vol. 24, pp. 49–79, 2005.

[38] Goldman, C. and Zilberstein, S., "Optimizing information exchange in cooperative multi-agent systems," in *Proceedings of International Joint Conference on Autonomous Agents and Multi Agent Systems*, 2003.

[39] Haller, H. and Lagunoff, R., "Genericity and Markovian behavior in stochastic games," *Econometrica*, vol. 68, no. 5, pp. 1231 – 1248, 2000.

[40] Hansen, E. A., "An improved policy iteration algorithm for partially observable MDPs," in *Advances in Neural Inform. Processing Systems (NIPS-97)*, vol. 10, pp. 1015–1021, MIT Press, Cambridge, MA, 1998.

[41] Hansen, E. A., Bernstein, D., and Zilberstein, S., "Dynamic programming for partially observable stochastic games," in *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, (San Jose, California), pp. 709 – 715, 2004.

[42] Hansen, E. A. and Zilberstein, S., "Lao*: A heuristic search algorithm that finds solutions with loops," *Artificial Intelligence*, vol. 129, pp. 35–62, 2001.

[43] Hernandez-Lerma, O. and Romera, R., "Multiobjective Markov control processes: a linear programming approach," *Morfismos*, vol. 8, no. 1, pp. 1–33, 2004.

[44] Howard, R., *Dynamic programming and Markov processes*. Cambridge, MA: MIT Press, 1960.

[45] Iyengar, G. N., "Robust dynamic programming," *Mathematics of Operations Research*, vol. 30, pp. 257–280, May 2005.

[46] Jia, Q.-S., "On state aggregation to approximate complex value functions in large-scale Markov decision processes," *IEEE Transactions on Automatic Control*, vol. 56, pp. 333–344, Feb. 2011.

[47] Kumar, A. and Zilberstein, S., "Dynamic programming approximations for partially observable stochastic games," in *Proceedings of the twenty-second international FLAIRS conference*, (Sanibel Island, Florida), pp. 547–552, 2009.

[48] Lewis, M. and Puterman, M., "A probabilistic analysis of bias optimality in unichain Markov decision processes," *IEEE Transactions on Automatic Control*, vol. 46, pp. 96–100, Jan. 2001.

[49] Li, L. and Sun, Z., "Dynamic energy control for energy efficiency improvement of sustainable manufacturing systems using Markov decision process," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 43, no. 5, pp. 1195 – 1205, 2013.

[50] Lin, Z., Bean, J. C., and White, C. C., "Genetic algorithm heuristics for finite horizon partially observed Markov decision problems," tech. rep., University of Michigan, Ann Arbor, MI 48109, USA, 1998.

[51] Lin, Z., Bean, J. C., and White, C. C., "A hybrid genetic/optimization algorithm for finite-horizon, partially observed Markov decision processes," *INFORMS Journal on Computing*, vol. 16, no. 1, pp. 27–38, 2004.

[52] Littman, M. L., "The Witness algorithm for solving partially observable Markov decision processes," tech. rep., Brown University, Department of Computer Science, 1994.

[53] Littman, M. L., *Algorithms for sequential decision making*. PhD thesis, Department of Computer Science, Brown University, Providence, Rhode Island 02912, 1996.

[54] Lovejoy, W. S., "Computationally feasible bounds for partially observed Markov decision processes," *Operations Research*, vol. 39, no. 1, pp. 162–175, 1991.

[55] Madani, O., Hanks, S., and Condon, A., "On the undecidability of probabilistic planning and infinite-hoziron partially observale decision problems," in *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, (Orlando), pp. 541 – 548, 1999.

[56] McMahan, H. and Gordon, G. J., "Fast exact planning in Markov decision processes," *Proceedings of the 15th International Conference on Automated Planninng and Scheduling*, 2005.

[57] Meyn, S. P., "The policy iteration algorithm for average reward Markov decision processes with general state space," *IEEE Transactions on Automatic Control*, vol. 42, no. 12, pp. 1663–1680, 1997.

[58] Monahan, G. E., "State of the art — a survey of partially observable Markov decision processes: theory, models, and algorithms," *Management Science*, vol. 28, no. 1, pp. 1–16, 1982.

[59] Moulik, S., Misra, S., Chakraborty, C., and Obaidat, M. S., "Prioritized payload tuning mechanism for wireless body area network-based healthcare systems," in *IEEE Global Communications Conference (GLOBECOM)*, pp. 2393 – 2398, 2014.

[60] Nair, R., Tambe, M., Roth, M., and Yokoo, M., "Communication for improving policy computation in distributed pomdps," in *Proceedings of International Joint Conference on Autonomous Agents and Multi Agent Systems*, 2004.

[61] Naser-Moghadasi, M., "A new graphical recursive pruning method for the incremental pruning algorithm," in *Advances in Artificial Intelligence, Lecture Notes in Computer Science*, no. 6437, pp. 232–242, 2010.

[62] Naser-Moghadasi, M., "Evaluating effect of two alternative filters for the incremental pruning algorithm on quality of pomdp exact solutions," *International Journal of Intelligence Science*, vol. 2, pp. 1–8, 2012.

[63] Nilim, A. and Ghaoui, L. E., "Robust control of Markov decision processes with uncertain transition matrices," *Operation Research*, vol. 53, no. 5, pp. 780–798, 2005.

[64] Oliehoek, F. A., "Decentralized POMDPs," in *Reinforcement Learning: State of the Art, Adaptation, Learning, and Optimization*, (Springer Berlin Heidelberg), 2012.

[65] Ortiz, O., Erera, A. L., and White, C. C., "State observation accuracy and finite-memory policy performance," *Operations Research Letters*, vol. 41, pp. 477 – 481, 2013.

[66] Papadimitriou, C. H. and Tsitsiklis, J. N., "The complexity of Markov decision processes," *Mathematics of Operations Research*, vol. 12, no. 3, pp. 441–450, 1987.

[67] Parthasarathy, T., "Existence of equilibrium stationary strategies in discounted stochastic games," *Sankya*, vol. 44, pp. 114–127, 1982.

[68] Puterman, M. L., *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994.

[69] Rabinovich, Z., Goldman, C. V., and Rosenschein, J. S., "The complexity of multiagent systems: the price of silence," in *Proceedings of the International Conference on Autonomous Agents and Multi Agent Systems*, (Melbourne, Australia), pp. 1102–1103, 2003.

[70] Rieder, U., *Equilibrium plans for nonzero sum Markov games, Game Theory and Related Topics*. North-Holland: ed. O. Moeschlin and D. Pallasche., 1979.

[71] Rogers, *Non-zero sum stochastic games*. PhD thesis, University of California at Berkeley, Berkeley, CA., 1969.

[72] Roijers, D. M., Vamplew, P., Whiteson, S., and Dazeley, R., "A survey of multi-objective sequential decision-making," *Journal of Artificial Intelligence Research*, vol. 48, pp. 67–113, 2013.

[73] Roth, M., Simmons, R., and Veloso, M., "Decentralized communication strategies for coordinated multi-agent policies," in *Multi-Robot Systems: From Swarms to Intelligent Automata*, vol. IV, Kluwer Avademic Publishers, 2005.

[74] Roy, B. V., "Performance loss bounds for approximate value iteration with state aggregation," *Mathematics of Operation Research*, vol. 31, no. 2, pp. 234–244, 2006.

[75] Saha, S., "Zero-sum stochastic games with partial information and average payoff," *Journal of Optimization Theory and Applications*, vol. 160, no. 1, pp. 344–354, 2014.

[76] Sanner, S., Goetschalckx, R., Driessens, K., and Shani, G., "Bayesian real-time dynamic programming," in *Proceedings of International Joint Conference on Aritifical Intelligence*, pp. 1784–1789, 2009.

[77] Seuken, S. and Zilberstein, S., "Formal models and algorithms for decentralized control of multiple agents," Tech. Rep. 05-68, Department of Computer Science, University of Massachusetts, Amherst, MA 01003, 2005.

[78] Shani, G., Pineau, J., and Kaplow, R., "A survey of point-based POMDP solvers," *Autonomous Agents and Multi-Agent Systems*, vol. 27, pp. 1–51, 2013.

[79] Shapley, L., "Stochastic games," *Proceedings of National Academy of Sciences of the United States of America*, vol. 39, pp. 1095 – 1100, 1953.

[80] Shlakhter, O., *Acceleration of iterative methods for Markov decision processes*. PhD thesis, Department of Mechanical and Industrial Engineering, University of Toronto, 2010.

[81] Smallwood, R. and Sondik, E., "The optimal control of partially observable Markov processes over a finite horizon," *Operations Research*, vol. 21, pp. 1071–1088, 1973.

[82] Sobel, M., "Non-cooperative stochastic games," *Annals of Mathematical Statistics*, vol. 42, pp. 1930 – 1935, 1971.

[83] Sondik, E. J., *The optimal control of partially observable Markov decision processes*. PhD thesis, Stanford University, Palo Alto, 1971.

[84] Sondik, E. J., "The optimal control of partially observable Markov processes over the infinite horizon: discounted costs," *Operations Research*, vol. 26, pp. 282–304, 1978.

[85] Sonu, E. and Doshi, P., "Generalized and bounded policy iteration for finitely nested interactive pomdps: Scaling up," in *12 th International Conference on Autonomous Agents and Multi-agent Systems (AAMAS)*, (Valencia, Spain), pp. 1039–1048, June 2012.

[86] Spaan, M., Gordon, G. J., and Vlassis, N., "Decentralized planning under uncertainty for teams of communicating agents," in *Proceedings of the International Joint Conference on Autonomous Agents and Multi Agent Systems*, pp. 249–256, 2006.

[87] Tan, C. H. and Hartman, J. C., "Sensitivity analysis in Markov decision processes with uncertain reward parameters," *Journal of Applied Probability*, vol. 48, pp. 954–967, 2011.

[88] Viswanathan, B., Aggarwal, V., and Nair, K., "Multiple criteria Markov decision processes," *Multiple Criteria Decision Making*, vol. 6, pp. 263–272, 1977.

[89] Wakuta, K., "Vector-valued Markov decision processes and the systems of linear inequalities," *Stochastic Processes and their Applications*, vol. 56, pp. 159–169, 1995.

[90] White, C., "A survey of solution techniques for the partially observed Markov decision process," *Annals of Operations Research*, vol. 32, pp. 215–230, 1991.

[91] White, C. and EL-Deib, H. K., "Parameter imprecision in finite state, finite action dynamic programs," *Operations Research*, vol. 34, no. 1, pp. 120–129, 1986.

[92] White, C. and EL-Deib, H. K., "Markov decision processes with imprecise transition probabilities," *Operations Research*, vol. 42, no. 4, pp. 739–749, 1994.

[93] White, C. and Kim, K., "Solution procedures for solving vector criterion Markov decision processes," *Large Scale System*, vol. 1, pp. 129–140, 1980.

[94] White, C. and Scherer, W., "Reward revision and the average reward Markov decision process," *OR Spektrum*, vol. 9, pp. 203–211, 1987.

[95] White, C., Thomas, L. C., and Scherer, W. T., "Reward revision for discounted Markov decision problems," *Operations Research*, vol. 33, no. 6, pp. 1299–1315, 1985.

[96] White, C. and White, D., "Markov decision process," *European Journal of Operational Research*, vol. 39, pp. 1–16, 1989.

[97] White, C. C. and Scherer, W. T., "Finite-memory suboptimal design for partially observed Markov decision processes," *Operations Research*, vol. 42, no. 3, pp. 439 – 455, 1994.

[98] Wiesemann, W., Kuhn, D., and Rustem, B., "Robust Markov decision processes," *Mathematics of Operation Research*, vol. 38, no. 1, pp. 153–183, 2013.

[99] Wingate, D. and Seppi, K. D., "Prioritization methods for accelerating mdp solvers," *Journal of Machine Learninnng Research*, vol. 6, pp. 851–881, 2005.

[100] Xuan, P., Lesser, V., and Zilberstein, S., "Communication decisions in multi-agent cooperation: Model and experiments," in *Proceedings of the Fifth International Conference on Autonomous Agents*, 2001.

[101] Yu, H., *Approximation solution methods for partially observable Markov and semi-Markov decision processes*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 2007.

[102] Zhang, H., "Partially observable Markov decision processes: A geometric technique and analysis," *Operations Research*, vol. 58, no. 1, pp. 214 – 228, 2010.

CHAPTER II

A LEADER-FOLLOWER PARTIALLY OBSERVED, MULTIOBJECTIVE MARKOV GAME

2.1 Introduction

The intent of this research is to generate a set of non-dominated policies from which one of two agents (the leader) with multiple objectives can select a most preferred policy to control a dynamic system that is also affected by the control decisions of the other agent (the follower). Such information can serve as input to a decision support system that, for example, is based on a deterministic version of multi-attribute utility theory (Keeney and Raiffa, 1993; Holloway and White, 2008). In this context, the results presented in this chapter generate options (i.e., policies) for consideration by (1) creating multiple generations of policies and eliminating all but the non-dominated set of policies from the last generation and (2) determining value scores for each of the policies in this non-dominated set.

The motivating application of this research is the operation of a liquid egg production facility in order to maximize the supply chain's productivity while minimizing its vulnerability to the intentional insertion of a biological or chemical toxin into the food production and distribution system (see Manning, et al., 2005; O'Ryan, et al., 1996; Sobel, et al., 2002). Liquid eggs are ingredients in a wide variety of foods, and contaminated liquid eggs could lead to significant morbidity and mortality. See Mohtadi and Murshid (2009) for background information about the importance of this application area. For this application, we assume the leader manages the production facility and is trying to balance two objectives: (1) maximize productivity and (2)

minimize vulnerability.

Although initially developed to model the liquid egg production process, the approach for decision aiding presented in this chapter can model a broad class of multi-epoch game applications when all agents are intelligent and adaptive. More generally, our objective is to provide decision support to the manager of a dynamic system when there is a second intelligent and adaptable agent with its own objective. This second agent can be cooperative, non-cooperative, or a mixture of both. Since most systems that we have been considering are in the private sector, we have assumed that the manager considers multiple objectives (e.g., maximize productivity, minimize risk) in operating the system. We believe that a multi-objective extension of the POMG represents a particularly appropriate model for this decision-making scenario.

Our approach to decision support is described as follows. We begin with an initial (i.e., first generation) set of possible leader policies. We then use a multi-objective genetic algorithm (MOGA) to create successive generations of presumably higher quality leader policies. We then determine the non-dominated set of the final generation of leader policies and present this set to the leader. The leader can then select the most preferred policy from this set for implementation.

The MOGA must know the fitness measures of each of the current leader policies before it can create the next generation of leader policies. The fitness measures of a leader policy are measures of how well the leader's objectives are met by this policy and depend on the follower's policy. We assume the follower knows the policy selected by the leader. This computationally advantageous assumption allows the POMG to be converted into a specially structured partially observed Markov decision process (POMDP) describing the follower's problem. Solving this POMDP determines the

follower policy. Given the leader policy and the follower policy, the fitness measures for the leader policy can then be computed by a value determination step.

We remark that the assumption that the follower knows the policy of the leader is a conservative assumption from the perspective of the leader and could unrealistically bias the game to the advantage of the follower. However, this bias is mollified by the fact that the leader and follower do not share the same data at each decision epoch, and hence the follower can only infer what action the leader will actually take. This assumption is also reasonable for many applications. For example, if the leader is a large governmental agency or corporation and the follower is an individual or group intent on attacking the leader, then it may be reasonable to assume that the follower will know more about the leader than the leader will know about the follower.

Many of the methodological characteristics of the decision support model presented in this chapter have been considered elsewhere in the decision, risk, and reliability analysis literatures. Models of intelligent agents or adversaries are examined by Cardoso and Diniz (2009). The single-period leader-follower (Stackelberg) game has been widely used to analyze the strategic interactions between two intelligent and adaptive agents. For example, Cavusoglu, et al., (2013) have studied the impacts of passenger profiling on airport security operations. Bakir (2011) analyzed resource allocation for cargo container transportation security. Other applications of the single-period leader-follower game are presented in Bier, et al., (2007, 2008) and Zhuang and Bier (2007). Multi-period games have been considered by Wang and Bier(2011), who examined a two-period leader-follower repeated game, and by Hausken and Zhuang (2011), who studied a multi-period game with myopic agents. Application of a completely observable stochastic game to overseas cargo container security can be found

in Bakir and Kardes (2009). Rothschild, et al., (2012) studied the imperfect observations of the other agent's action by a k-level game theory model.

Each of above methodological characteristics is intended to enhance the realism of the respective model. Our model extends the existing literature on stochastic games by explicitly considering the multi-period interaction of two non-myopic agents, a leader and a follower, each of whom adjust its decisions according to the other agent's decisions over an infinite planning horizon for the case where neither agent has complete information about the other agent. This model can be used for cooperative games, non-cooperative games, or a mixture of both. Furthermore, we introduce multi-objective optimization to the general-sum partially observed Markov game to enable multi-attribute decision making. By combining these characteristics into a single model, as has been done in this chapter, we believe that the modeling realism of the resulting model has been further enhanced. However, and not surprisingly, additional model realism has resulted in increased computational challenges. Dealing with these challenges is the focus of much of this chapter.

We summarize the contributions of this chapter as follows. Based on the POMG, POMDP, and MOGA models and computational results associated with the latter two models, we have identified a set of assumptions and developed results that have created a pathway to a tractable heuristic for the POMG. The set of assumptions and results are:

- At the policy level (before the game begins), we assume the follower becomes fully aware of the policy selected by the leader. This assumption allows the POMG to be transformed into a more computationally familiar POMDP.
- We assume the leader's policy is a finite-memory policy. This assumption guarantees that the resulting specially structured POMDP has a computationally

useful epoch-invariant, finite dimensional sufficient statistic (Proposition 2.1).

- The resulting follower policy is a perfect memory policy, which we approximate with a finite-memory policy. This finite-memory approximation guarantees that the value determination step necessary to compute the fitness measures for the given leader policy has an epoch-invariant, finite dimensional sufficient statistic (Proposition 2.2).

- We then are able to compute the fitness measures for each current generation leader policy. The MOGA uses these fitness measures to create the next generation of leader policies.

- The non-dominated set of the final generation of leader policies created by the MOGA and the fitness measures of each non-dominated leader policy then represents the output of our process.

The chapter is organized as follows. We review the pertinent literature associated with the MOGA, POSG, and POMDP in Section 2.2. In Section 2.3, we describe the MOGA in more detail, show how the fitness measures are computed using the POMG and the POMDP, and present equilibrium conditions. In Section 2.4 we illustrate how our decision support procedure can be applied to a simplified liquid eggs supply chain security problem and demonstrate computational feasibility. Section 2.5 summarizes research results and discusses future research directions.

2.2 Literature Review

The research presented in this chapter contributes to the POSG literature by proposing a framework for a general sum leader-follower partially observed Markov game with multiple criteria for the leader and a corresponding heuristic solution procedure. The proposed solution procedure uses solution techniques from the existing POMDP

and MOGA literature. We now review the pertinent POMDP, MOGA and POSG literature.

2.2.1 The partially observed Markov decision process

The POMDP is a model of single agent sequential decision making under uncertainty that takes into consideration possibly inaccurate and/or costly observations of the state of the system under control. Relative to the completely observed Markov decision process (i.e., the MDP; see Puterman, 1994), the POMDP represents a more general but significantly more computationally challenging model. In seminal research, Smallwood and Sondik (1973) and Sondik (1978) showed that the optimal cost function of the POMDP has a computationally interesting structure and presented successive approximations approaches for solving the POMDP that exploited this structure. Zhang (2010) revisited these structural results and convergence properties by exploiting the dual relationship between hyperplanes and points in the POMDP and related the solution of the POMDP as a Minkowski sum problem in computational geometry. Monahan (1982), Eagle (1984), and White and Scherer (1989) presented improved algorithms based on these structural results. Detailed descriptions of other exact algorithms can be found in Cheng (1988), Littman (1994a), Cassandra, et al., (1994), Cassandra, et al., (1997), Feng and Zilberstein (2004), Lin, et al., (1998, 2004) and Naser-Moghadasi (2012). Surveys of related solution techniques and complexity analyses for the POMDP can be found in Monahan (1982), Lovejoy (1991a), White (1991), Cassandra (1994) and Poupart (2005). The solution technique in Lin, et al., (1998, 2004) is used to solve the follower's optimization problem in this chapter, for any given leader's policy.

In the development of approximate solution techniques for POMDPs, point-based

value iteration (PBVI) was presented and analysed in Pineau, et al., (2003) and Shani, et al., (2013). Platzman (1977, 1980), White and Scherer (1994), Littman (1994b), Hauskrecht (1997), Hansen (1998a, 1998b), Poupart and Boutilier (2004), Poupart (2005) examined finite memory policies and finite-state controllers. The concept of a finite memory policy is used in this chapter to approximate the follower's best response perfect memory policy in order to calculate the leader's value function. Other approximate methods for POMDPs are reviewed in Hauskrecht (2000), Aberdeen (2003), and Yu (2007).

2.2.2 The partially observable stochastic game

The stochastic game introduced by Shapley (1953) represents a multi-agent planning problem in a stochastic environment. In this setting, each player considers the consequences of its own action and the actions that its opponents or teammates may take. Algorithms for computing Nash equilibria for stochastic games can be found in Raghavan and Filar (1991), Filar, et al., (1997), and Ummels (2010). Two elements in our partially observed Markov game model, Stackelberg equilibrium and multi-objective decision making, have been studied in completely observed stochastic games. For example, Vorobeychik and Singh (2012), Vorobeychik, et al., (2012, 2014), Letchford, et al., (2012) have introduced Stackelberg equilibria to stochastic games. Canu and Mouaddib (2011) investigated how to coordinate a group of robots for planet exploration by a vector-valued stochastic game with Stackelberg equilibrium. This chapter further extends the Stackelberg equilibrium concept and multi-objective optimization to a general-sum partially observed Markov game.

The POSG is a new, relatively unexamined generalization of the stochastic game, where the states of the game are not precisely observed by the players and all players

make decisions based on these partial observations. Although POSGs provide a robust framework for multi-agent planning, Bernstein, et al., (2002) showed that POSGs are computationally intractable when problem size grows. Rabinovich, et al., (2003) has shown that even epsilon-optimal approximations are NP-hard. As a result, most of the work in the literature has focused on POSGs with special structure. McEneaney (2004) focused on a game where only one player has imperfect information. Ghosh, et al., (2004), Oliehoek, et al., (2005), and Bopardikar and Hespanha (2011) studied a zero-sum version of the POSG. Emery-Montemerlo, et al., (2004) approximated a cooperative POSG by a series of Bayesian games. Another cooperative POSG, called a decentralized partially observable Markov decision process (DEC-POMDP), has also been extensively studied by Becker, et al., (2004), Bernstein, et al., (2005), Seuken and Zilberstein (2007), and Oliehoek, et al., (2008). A survey of the DEC-POMDP can be found in Oliehoek (2012). A cooperative POSG is applicable only when the players are strictly cooperative. In contrast, this chapter focuses on a generalized solution procedure that can be used to solve general-sum multi-objective partially observed Markov games when the players are cooperative, noncooperative, or a mixture of both.

For the general-sum partially observed Markov game, Hespanha and Prandini (2001) proved the existence of a Nash equilibrium in a two-player finite-horizon problem. Hansen, et al., (2004) developed a dynamic program for general POSGs by pruning very weakly dominated strategies and then showed that this dynamic programming approach can achieve optimality for cooperative settings. However, this approach is computationally intractable for all but the smallest problems. Kumar and Zilberstein (2009) developed an approximate solution procedure for the POSG based on Hansen's work. While Hansen and Kumar's work focuses on the Nash equilibrium, this chapter is interested in Stackelberg equilibrium that can be used in security

applications. Interactive POMDPs (I-POMDPs) addressed in Gmytrasiewicz and Doshi (2005) demonstrated another framework for general multi-agent planning by augmenting the state space to include models of other players' behaviour. However, these I-POMDPs are difficult to solve optimally because the states are infinitely nested by construction (Doshi, 2012). The method we propose in this chapter considers the case where the players make their decisions based on their past histories. Hence we avoid the infinitely nested modeling, which may make I-POMDPs difficult to solve in general (Doshi, 2012). In addition, connected with multi-objective optimization, our framework can also be used for the agents with multiple criteria.

2.2.3 The multi-objective genetic algorithm

Genetic algorithms, introduced by Holland (1975), are adaptive heuristic search techniques that mimic the process of natural evolution. A genetic algorithm represents each feasible problem solution in a population of solutions as a genome or chromosome and begins with an initial population of feasible solutions. Solutions having high measures of fitness are preferably selected during each generation to produce the next generation of solutions, which typically have improved fitness measures due to the application of genetic (e.g., mutation and crossover) operators. After a number of generations, the population presumably evolves to optimal or near-optimal solutions. Goldberg (1989), Forrest (1993) and Srinivas and Patnaik (1994) present surveys of genetic algorithms and their related theories.

Multi-objective genetic algorithms (MOGA) are designed for the simultaneous optimization of multiple, often competing objectives. Usually the optimal solutions are a set of points, called the Pareto-optimal set, in the sense that no improvement can be made in any objective without sacrificing the other objectives. The MOGA pushes the

Pareto frontier towards the ideal optimal set of solutions as the algorithm proceeds. MOGA algorithms include: the vector evaluated GA (VEGA) (Schaffer, 1985), the Niched Pareto GA (NPGA) (Horn, et al., 1994), the Pareto Envelope-based Selection Algorithms (PESA) (Corne, et al., 2000) and the Fast Non-dominated sorting GA (NSGA-II) (Deb, et al., 2002). Surveys are presented in Coello (2000) and Konak, et al.,(2006). MOGAs have been widely applied in optimization and decision making problems (see Ponnambalam et al., 2001; Deb, 2001; Ombuki, et al., 2006; Lin and Gen, 2008; Bowman, et al., 2010; Yildirim and Mouzon 2012). This chapter will use a MOGA, the NSGA-II, to generate policies from which the leader will choose a most preferred policy.

2.3 Model and Analysis

We present the POMG model in Section 2.3.1. In order to determine an optimal response policy for the follower, the POMDP is constructed by combining the POMG model with the leader's policy that is currently under consideration. The resulting POMDP is presented and examined in Section 2.3.2. For computational reasons for the leader, we require that the leader and follower policies to be finite memory policies. However, the POMDP constructs a perfect memory follower policy. In Section 2.3.3, we present an approach for determining a finite-memory approximation of a perfect memory policy. In order for the MOGA to determine the next generation of leader policies, fitness measures must be calculated for each policy in the current generation of leader policies. Each fitness measure is a measure of an objective of the leader. In Section 2.3.4 we present an approach for determining the fitness measures, for any given leader policy and follower policy. Section 2.3.5 describes how we used a MOGA to generate a non-dominated set of leader policies. Section 2.3.6 addresses equilibria.

2.3.1 Partially Observed Markov Game

The partially observed Markov game (POMG) serves as the modeling basis of our decision support system design. The POMG is comprised of:

Decision epochs: Let $\{0, 1, \ldots\}$ be the set of all decision epochs when both agents select actions simultaneously. Thus, the problem horizon is countable and infinite.

State spaces: Let S^L and S^F be the state spaces of the leader and the follower, respectively. Both spaces are epoch-invariant and finite. At decision epoch t, let $s^L(t)$ be the leader's state, $s^F(t)$ be the follower's state, and denote $s(t) = \{s^L(t), s^F(t)\}$.

Action spaces: Let A^L and A^F be the epoch-invariant action spaces of the leader and the follower, both of which are finite. At decision epoch t, let $a^L(t)$ be the leader's action, $a^F(t)$ be the follower's action, and denote $a(t) = \{a^L(t), a^F(t)\}$.

Observation spaces: Let Z^L and Z^F be the observation spaces of the leader and the follower, both of which are epoch-invariant and finite. At decision epoch t, let $z^F(t)$ be the follower's observation of the leader's state, $z^L(t)$ the leader's observation of the follower's state, and denote $z(t) = \{z^L(t), z^F(t)\}$.

Systems dynamics: We assume the epoch-invariant probability $P(z(t+1), s(t+1)|s(t), a(t))$ is given. Note that

$$P(z(t+1), s(t+1)|s(t), a(t)) = P(z(t+1)|s(t+1), s(t), a(t))P(s(t+1)|s(t), a(t)),$$

where $P(z(t+1)|s(t+1), s(t), a(t))$ is referred to as the state observation probability and $P(s(t+1)|s(t), a(t))$ is referred to as the state transition probability.

Single Period Cost and Criteria: Let $c^F(s(t), a(t))$ be the decision epoch invariant single period cost accrued by the follower at epoch t, given $s(t)$ and $a(t)$, and let $c_i^L(s(t), a(t))$ be the decision epoch invariant single period cost accrued by the leader with respect to criterion i at epoch t, given $s(t)$ and $a(t)$. The criteria under consideration are the concomitant expected total discounted costs over the infinite horizon.

Policies: A policy π^k for agent $k \in \{L = \text{leader}, F = \text{follower}\}$ is a mapping from what agent k knows at epoch t into its set of available actions, A^k. Policies can be random and hence described by conditional probabilities. We restrict our interest to stationary policies. Stationary policies tend to be easy to implement and in many cases, e.g., the determination of optimal follower response policies for a broad class of scalar criteria, sufficiently rich to contain an optimal policy.

Information patterns: We assume the leader chooses a finite-memory policy π^L, a policy which is a mapping from the set of all $\mathscr{I}^L(t, \tau)$ into A^L at decision epoch t, where for $k \in \{L, F\}$, $\mathscr{I}^k(t, \tau) = \{z^k(t), \ldots, z^k(t - \tau + 1), s^k(t), \ldots, s^k(t - \tau + 1), a^k(t - 1), \ldots, a^k(t - \tau)\}$. Thus, π^L is of the form $\{P(a^L(t)|\mathscr{I}^L(t, \tau))\}$. Hence, when selecting an action, we assume the leader knows the current and τ most recent observations of the other agent's state, its current and τ most recent state values, and the τ most recent actions it has selected. These are reasonable assumptions for many applications. For example, in the context of the application mentioned in Section 2.4, the defender may know its defensive resource allocation, but have little knowledge about the attacker's status. As another example, a company knows its own state, but it may know little about its competitor's state. Below we will assume the leader knows information in addition to $\mathscr{I}^L(t, \tau)$ in order to determine its criteria values.

Given π^L, the POMG becomes a specially structured POMDP for the follower, which the follower solves in order to determine an optimal *perfect-memory* response policy; that is, this response policy is a mapping into A^F from the set of all $\mathscr{I}^F(0) = \{s^F(0), y^F(0)\}$ when $t = 0$ and from the set of all $\mathscr{I}^F(t) = \{z^F(t), \ldots, z^F(1), s^F(t), \ldots, s^F(0), a^F(t-1), \ldots, a^F(0), y^F(0)\}$ when $t \geq 1$, where for $m, k \in \{L, F\}$, $m \neq k$, $y^k(0) = \{P(\mathscr{I}^m(0, \tau))\}$. Let $y^k(t) = \{P(\mathscr{I}^m(t, \tau)|\mathscr{I}^k(t))\}$ when $t \geq 1$, where $y^k(t)$ is a "belief" array that indicates what agent k can infer about the other agent's information pattern, i.e., $\mathscr{I}^m(t, \tau), m \neq k$. We will show later that $\{s^F(t), y^F(t)\}$ is a sufficient statistic for this POMDP, given this perfect-memory information pattern. We remark that the arrays $\{y^F(t)\}$ are needed as part of the sufficient statistic in order for the follower to infer the leader's action selection in determining the follower's criterion value. We further remark that Bayes' Rule can be used to compute $y^k(t+1)$, given $y^k(t)$, $z^k(t+1)$, $s^k(t+1)$, and $a^k(t)$: Let $\varsigma^k(t) = \{z^k(t), s^k(t), a^k(t-1)\}$ and $\varsigma(t) = \{\varsigma^L(t), \varsigma^F(t)\}$. Without loss of generality, we determine $y^F(t+1)$, given $y^F(t)$ and $\varsigma^F(t+1)$. Note, $\mathscr{I}^L(t+1, \tau) = \{\varsigma^L(t+1), \mathscr{I}^L(t, \tau-1)\}$ and $\mathscr{I}^F(t+1) = \{\varsigma^F(t+1), \mathscr{I}^F(t)\}$. Then,

$$P(\varsigma^L(t+1), \mathscr{I}^L(t, \tau-1)|\varsigma^F(t+1), \mathscr{I}^F(t)) = \sum_{\varsigma'} P(\varsigma^L(t+1), \mathscr{I}^L(t, \tau)|\varsigma^F(t+1), \mathscr{I}^F(t)),$$

where $\varsigma' = \varsigma^L(t - \tau + 1)$.

Note

$$P(\varsigma(t+1), \mathscr{I}^L(t, \tau)|\mathscr{I}^F(t)) = P(\varsigma^L(t+1), \mathscr{I}^L(t, \tau)|\varsigma^F(t+1), \mathscr{I}^F(t))P(\varsigma^F(t+1)|\mathscr{I}^F(t))$$

and that

$$P(\varsigma^F(t+1)|\mathscr{I}^F(t)) = \sum_{\varsigma''} \sum_{\mathscr{I}} P(\varsigma(t+1), \mathscr{I}^L(t, \tau)|\mathscr{I}^F(t)),$$

where $\varsigma'' = \varsigma^L(t+1), \mathscr{I} = \mathscr{I}^L(t, \tau)$.

Now,

$$P(\varsigma(t+1), \mathscr{I}^L(t, \tau)|\mathscr{I}^F(t)) = P(\varsigma(t+1)|\mathscr{I}^L(t, \tau), \mathscr{I}^F(t))P(\mathscr{I}^L(t, \tau)|\mathscr{I}^F(t)).$$

Then,

$P(\varsigma(t{+}1)|\mathscr{I}^L(t,\tau),\mathscr{I}^F(t)) = P(z(t{+}1),s(t{+}1)|a(t),\mathscr{I}^L(t,\tau),\mathscr{I}^F(t))P(a(t)|\mathscr{I}^L(t,\tau),\mathscr{I}^F(t))$.

Thus, we note that $P(\varsigma^L(t{+}1),\mathscr{I}^L(t,\tau{-}1)|\varsigma^F(t{+}1),\mathscr{I}^F(t))$ is a function of $\{P(\mathscr{I}^L(t,\tau)|\mathscr{I}^F(t))\}$.

The reason for requiring the leader's policy to be finite-memory now becomes clear: to guarantee that there is a sufficient statistic for this POMDP that has a computationally desirable finite and t-invariant number of elements, whereas the cardinality of $\mathscr{I}^F(t)$ increases as t increases. A similar POMDP construction can be found in Kandori and Obara (2010) for a repeated game, whereas we construct POMDP to solve a partially observed stochastic game.

The criterion value for each of the leader's objectives (i.e. fitness measures) is well defined, given leader and follower policies. However, as is true for the follower's POMDP, we wish to determine a sufficient statistic that has a finite and t-invariant number of elements when determining criteria values. We therefore assume that the follower's policy π^F is a finite-memory policy. An approximation procedure for determining a finite-memory policy π^F (i.e., a policy of the form $\{P(a^F(t)|\mathscr{I}^F(t,\tau))\}$) from the optimal perfect-memory policy determined from the follower's POMDP (which is of the form $\{P(a^F(t)|\mathscr{I}^F(t))\}$) is given later in the chapter.

We assume that the data available to the leader for value determination are $\mathscr{I}^L(0) = \{\mathscr{I}^L(0,\tau), y^L(0)\}$ when $t = 0$ and are $\mathscr{I}^L(t) = \{z^L(t),\ldots,z^L(1),s^L(t),\ldots,s^L(1),a^L(t{-}1),\ldots,a^L(0),\mathscr{I}^L(0,\tau),y^L(0)\}$ when $t \geq 1$. We will show later that $\{\mathscr{I}^L(t,\tau),y^L(t)\}$ is a sufficient statistic for this value determination step, given this perfect-memory information pattern. We note that this sufficient statistic also has a finite and t-invariant number of elements.

In summary, we assume:

- The leader knows $\mathscr{I}^L(t,\tau)$ at decision epoch t for selecting its action $a^L(t)$.
- Given the leader's policy, the follower determines its optimal perfect-memory response policy, assuming it knows $\mathscr{I}^F(t)$ at decision epoch t for selecting its action $a^F(t)$ and for inferring the selection of the leader's action $a^L(t)$.
- Given the follower's perfect-memory best response policy, the leader approximates the follower's response policy with a finite-memory policy that depends on $\mathscr{I}^F(t,\tau)$ at decision epoch t for selecting the follower's action $a^F(t)$.
- The leader then takes its policy and the finite-memory approximation to the follower's policy in order to determine the leader's criteria values, assuming the leader knows $\mathscr{I}^L(t)$ at decision epoch t for selecting its action $a^L(t)$ and for inferring the selection of the follower's action $a^F(t)$.

We note that the sufficient conditions for perfect-memory information patterns $\mathscr{I}^F(t)$ and $\mathscr{I}^L(t)$ are asymmetric, which is a result of the facts that:

- In determining $v^F(\mathscr{I}^F(t))$ and an optimal perfect-memory policy for the follower, we only need to provide a priori data to support being able to infer the data needs of the leader's finite memory policy.
- In determining the criteria values for the leader, value determination requires consideration of both the leader's policy and the follower's policy, which requires a priori data to support the leader's policy and a priori data to be able to infer the data needs of the finite-memory approximation to the follower's optimal policy.

We remark that the POMDP presented in (Smallwood and Sondik, 1973) and elsewhere typically assumes the decision maker's state is partially observed, whereas we

are assuming each agent completely observes its state and partially observes the state of the other agent. In the context of the POMDP, the information patterns that are assumed in this chapter are analogous to the special case where the underlying state of the POMDP is a two-vector, one element of which is completely observed and the other element is partially observed.

Objectives: The follower's objective is to select a stationary policy that optimizes its criterion (or in reality we settle for an ϵ-optimal policy if an optimal policy is difficult or impossible to obtain). The leader's objective is to optimize all criteria under consideration in some balanced manner, with this balance being determined by the leader. Our objective is to provide the leader with a non-dominated set of policies from which to choose a single policy for implementation.

2.3.2 Determination of a Best Response Policy $\bar{\pi}^F$, Given a leader policy π^L

Given the leader's policy π^L, let $v^F(\mathscr{I}^F(t))$ be the follower's optimal criterion value, given $\mathscr{I}^F(t)$. Then, according to results in (Puterman, 1994; Chapter 6), v^F uniquely satisfies

$$v^F = H^F v^F \tag{2.1}$$

where for any v,

$$[H^F v](\mathscr{I}^F(t)) = \min_{a^F(t)} h^F(\mathscr{I}^F(t), a^F(t), v),$$

$$h^F(\mathscr{I}^F(t), a^F(t), v) = E\{c^F(s(t), a(t)) + \beta v(\mathscr{I}^F(t+1)) | \mathscr{I}^F(t), a^F(t)\},$$

and where E is the expectation operator. Further, a policy that attains the above minimum for all $\mathscr{I}^F(t)$ is an optimal perfect memory policy. We now state our first result.

Proposition 2.1. *Assume π^L is given. Then for each $\mathscr{I}^F(t)$ there is an at most countable set of arrays $\Gamma^*(s^F(t))$ that only depends on $s^F(t)$, such that:*

$$v^F(\mathscr{I}^F(t)) = \min\{\sum \gamma(\mathscr{I}^L(t,\tau))P(\mathscr{I}^L(t,\tau)|\mathscr{I}^F(t)) : \gamma \in \Gamma^*(s^F(t))\},$$

where the sum is over all $\mathscr{I}^L(t,\tau)$.

Proof. Assume v and Γ are such that

$$v(\mathscr{I}^F(t)) = \min\{\sum \gamma(\mathscr{I}^L(t,\tau))P(\mathscr{I}^L(t,\tau)|\mathscr{I}^F(t)) : \gamma \in \Gamma(s^F(t))\},$$

where the sum is over all $I^L(t,\tau)$. Then the analysis, following the same line of arguments in (Smallwood and Sondik, 1973), shows that

$$h^F(\mathscr{I}^F(t), a^F(t), v) = \min\{\sum \gamma'(\mathscr{I}^L(t,\tau))P(\mathscr{I}^L(t,\tau)|\mathscr{I}^F(t)) : \gamma' \in \Gamma'(s^F(t), a^F(t))\},$$

where the sum is over all $\mathscr{I}^L(t,\tau)$. If $\gamma' \in \Gamma'(s^F(t), a^F(t))$ then γ' is of the form

$$\gamma(\mathscr{I}^L(t,\tau)) = \sum_{a^L(t)} P(a^L(t)|\mathscr{I}^L(t,\tau))[c^F(s(t), a(t))$$

$$+ \beta \sum_{s(t+1)} \sum_{z(t+1)} \gamma^{i,j}(z^L(t+1), s^L(t+1), a^L(t), \mathscr{I}^L(t,\tau-1))P(z(t+1), s(t+1)|s(t), a(t))],$$

where $\gamma^{i,j}$ can be any element in $\Gamma(s^F(t+1))$ for each $s^F(t+1) = i$ and $z^F(t+1) = j$. And $\{z^L(t+1), s^L(t+1), a^L(t), \mathscr{I}^L(t,\tau-1)\} = \mathscr{I}^L(t+1,\tau)$. Then,

$$[H^F v](\mathscr{I}^F(t)) = \min\{\sum_{\mathscr{I}^L(t,\tau)} \gamma''(\mathscr{I}^L(t,\tau))P(\mathscr{I}^L(t,\tau)|I^F(t)) : \gamma'' \in \Gamma''(s^F(t))\},$$

where $\Gamma''(s^F(t)) = \cup_{a^F(t)}\Gamma'(s^F(t), a^F(t))$.

The operator H^F is a contraction operator on the Banach space comprised of all functions mapping $\mathscr{I}^F(t)$ into the real line, having as its norm the supremum norm, and as a result, the sequence $\{v^n\}$, where $v^{n+1} = H^F v^n$, converges to v^F for any given v^0. The above result indicates that H^F preserves piecewise linearity and concavity and in the limit preserves concavity. □

With regard to the implications of Proposition 2.1 and results in (Puterman, 1994; Chapter 6), v^F and hence an optimal policy can depend on $\mathscr{I}^F(t)$ only through $(s^F(t), y^F(t))$. Hence, $(s^F(t), y^F(t))$ is a sufficient statistic. Further, v^F is concave in $y^F(t)$. Additionally, if $\Gamma^*(s^F(t))$ is a finite set of arrays for all $s^F(t)$, then v^F is piecewise linear. Note that the dimension of $(s^F(t), y^F(t))$ is finite and t-invariant. Note further that the finite dimensionality of $y^F(t)$ follows directly from the finite-memory assumption imposed on π^L. Thus, assuming $\Gamma^*(s^F(t))$ is a finite set of arrays and v^F is described in terms of $(s^F(t), y^F(t))$, v^F has a finite representation. We remark that the cardinality of $\Gamma''(s^F(t))$ can be substantially larger than the cardinality of $\Gamma(s^F(t))$, where both $\Gamma(s^F(t))$ and $\Gamma''(s^F(t))$ are defined in the proof of Proposition 2.1. Techniques for reducing the cardinality of $\Gamma''(s^F(t))$ can be found in White (1991) and Lin, et al., (1998, 2004).

2.3.3 A Finite-Memory Approximation to $\bar{\pi}^F$

As noted above, an optimal policy $\bar{\pi}^F$ for the follower that achieves the minimum in Equation (2.1) depends on $\mathscr{I}^F(t)$ and hence $\bar{\pi}^F$ is a perfect-memory policy $\bar{\pi}^F : \{\mathscr{I}^F(t)\} \rightarrow A^F$. In order to ensure that the leader's criteria have finite representation, the follower policy must also be a finite-memory policy from the leader's perspective. There are a variety of ways to approximate a perfect-memory policy by a finite-memory policy. We determine a finite-memory (approximate) policy from a given perfect memory policy as follows. We note since $(s^F(t), y^F(t))$ is a sufficient statistic, the perfect-memory policy $\bar{\pi}^F$ is a mapping $\bar{\pi}^F : \{s^F(t), y^F(t)\} \rightarrow A^F$. We also observe that $\{(\mathscr{I}^F(t,\tau), y^F(t-\tau)), t = 1, 2, \ldots\}$ is also a sufficient statistic for this problem, with $y^F(t-\tau)$ representing the influence of data determined up through epoch $t-\tau$, by noting $y^F(t)$ can be determined from repeated application of Bayes'

rule, given $\mathscr{I}^F(t,\tau), y^F(t-\tau)$, and π^L. By (arbitrarily) assuming a uniform distribution over $y^F(t-\tau)$, we can determine a finite-memory approximate policy in the form of $\{P(a^F(t)|\mathscr{I}^F(t,\tau))\}$. The quality of a finite memory policy has been investigated in White and Scherer(1994). Although we have not done extensive numerical testing, we have observed that finite memory approximations of optimal perfect memory policies can be quite good, even for small τ. Below is an example from our numerical experiment with 2 stats, 2 observations and 2 actions for each agent. The resulting γ vectors for $s^F = 1$ are:

(s^L, z^L)	(s_1, z_1)	(s_1, z_2)	(s_2, z_1)	(s_2, z_2)
$a^F = a_1$	[4.4784	4.1677	4.4784	4.1677]
$a^F = a_2$	[4.7137	2.5929	4.7137	2.5929]

At the related information pattern $(s^F = 1, z^F = 1)$, the follower will select action a_2 for $y^F(t) = P(s^L(t), z^L(t)|\mathscr{I}^F(t))$ where $P(s_1, z_1|\mathscr{I}^F(t)) + P(s_2, z_1|\mathscr{I}^F(t)) \geq 0.85$. Let $\tau = 1$. Drawing 500 samples of $y^F(t-1)$ from a uniform distribution over $S^L \times Z^L$ indicates that the resulting approximate finite memory policy is $P(a^F(t) = a_2|\mathscr{I}^F(t, \tau = 1)) = 0.938$. .

In the following context, we use π^F to denote the *finite-memory* response policy approximation to the optimal perfect-memory policy for the follower. Other approaches for determining a finite-memory approximation are under consideration and are topics for future research.

2.3.4 Fitness Measure Determination

Let $v_i^L(\pi^L, \pi^F; \mathscr{I}^L(t))$ be the criterion value for the leader's i^{th} criterion, given $\mathscr{I}^L(t)$, a leader policy $\pi^L = \{P(a^L(t)|\mathscr{I}^L(t,\tau))\}$, and a follower policy $\pi^F = \{P(a^F(t)|\mathscr{I}^F(t,\tau))\}$. For notational simplicity, we assume that the dependence of $v_i^L(\pi^L, \pi^F; \mathscr{I}^L(t))$ on

(π^L, π^F) is implicit; hence, $v_i^L(\mathscr{I}^L(t)) = v_i^L(\pi^L, \pi^F; \mathscr{I}^L(t))$. Then according to results in (Puterman, 1994; Chapter 6), v_i^L uniquely satisfies

$$v_i^L(\mathscr{I}^L(t)) = h_i^L(\mathscr{I}^L(t), v_i^L)$$

for all $\mathscr{I}^L(t)$, where

$$h_i^L(\mathscr{I}^L(t), v) = E\{c_i^L(s(t), a(t)) + \beta v(\mathscr{I}^L(t+1))|\mathscr{I}^L(t)\}.$$

We now show that $v_i^L(\mathscr{I}^L(t))$ is dependent on $\mathscr{I}^L(t)$ only though $(\mathscr{I}^L(t,\tau), y^L(t))$. Thus, $(\mathscr{I}^L(t,\tau), y^L(t))$ is a sufficient statistic.

Proposition 2.2. *Assume (π^L, π^F) are given finite-memory policies. Then, there is a function g_i^* such that*

$$v_i^L(\mathscr{I}^L(t)) = \sum g_i^*(\mathscr{I}^L(t,\tau), \mathscr{I}^F(t,\tau))P(\mathscr{I}^F(t,\tau)|\mathscr{I}^L(t)),$$

where the sum is over all $\mathscr{I}^F(t,\tau)$. Further, g_i^* is the unique solution of the equation

$$g_i^*(\mathscr{I}^L(t,\tau), \mathscr{I}^F(t,\tau)) = \sum{}^1 P(s(t), a(t)|\mathscr{I}^F(t,\tau), \mathscr{I}^L(t,\tau))\{c_i^L(s(t), a(t))$$
$$+ \beta \sum{}^2 g_i^*[(\mathfrak{z}^L(t+1), \mathscr{I}^L(t,\tau-1)), (\mathfrak{z}^F(t+1), \mathscr{I}^F(t,\tau-1))]P(z(t+1), s(t+1)|s(t), a(t))\},$$

where $\mathfrak{z}^k(t) = \{z^k(t), s^k(t), a^k(t-1)\}$, $\sum^1$ is over all $s(t)$ and $a(t)$, and $\sum^2$ is over all $z(t+1)$ and $s(t+1)$.

Proof. We remark that since (π^L, π^F) is assumed given, $P(s(t), a(t)|\mathscr{I}^F(t,\tau), \mathscr{I}^L(t,\tau))$ is well defined. Assume there is a function g such that

$$v(\mathscr{I}^L(t)) = \sum g(\mathscr{I}^L(t,\tau), \mathscr{I}^F(t,\tau))P(\mathscr{I}^F(t,\tau)|\mathscr{I}^L(t)),$$

where the sum is over all $\mathscr{I}^F(t,\tau)$. Then it is straightforward to show that there is a function g' such that

$$h_i^L(\mathscr{I}^L(t), v) = \sum g'(\mathscr{I}^L(t,\tau), \mathscr{I}^F(t,\tau))P(\mathscr{I}^F(t,\tau)|\mathscr{I}^L(t)),$$

where the sum is over all $\mathscr{I}^F(t,\tau)$, and

$$g'(\mathscr{I}^L(t,\tau),\mathscr{I}^F(t,\tau)) = \sum\nolimits^1 P(s(t),a(t)|\mathscr{I}^F(t,\tau),\mathscr{I}^L(t,\tau))\{c_i^L(s(t),a(t))$$
$$+\beta\sum\nolimits^2 g[(\mathfrak{z}^L(t+1),\mathscr{I}^L(t,\tau-1)),(\mathfrak{z}^F(t+1),\mathscr{I}^F(t,\tau-1))]P(z(t+1),s(t+1)|s(t),a(t))\},$$

and where $\mathfrak{z}^k(t) = \{z^k(t), s^k(t), a^k(t-1)\}$, $\sum^1$ is over all $s(t)$ and $a(t)$, and $\sum^2$ is over all $z(t+1)$ and $s(t+1)$.The result follows directly from the following facts:

- The operator H^L, where $[H^L v](\mathscr{I}^L(t)) = h_i^L(\mathscr{I}^L(t), v)$, is a contraction operator on the Banach space comprised of all functions mapping the set of all $\mathscr{I}^L(t)$ into the real line, having as its norm the supremum norm.

- As a result, the sequence $\{v^n\}$, where $v^{n+1} = H^L v^n$, converges to v^L for any given v^0.

Determine $y^k(t+1)$, given $y^k(t), z^k(t+1), s^k(t+1)$ and $a^k(t)$: Let $\varsigma^k(t) = \{z^k(t), s^k(t), a^k(t-1)\}$ and $\varsigma(t) = \{\varsigma^L(t), \varsigma^F(t)\}$. Without loss of generality, we determine $y^F(t+1)$, given $y^F(t)$ and $\varsigma^F(t+1)$. Note,
$\mathscr{I}^L(t+1,\tau) = \{\varsigma^L(t+1), \mathscr{I}^L(t,\tau-1)\}$ and $\mathscr{I}^F(t+1) = \{\varsigma^F(t+1), \mathscr{I}^F(t)\}$. Then,

$$P(\varsigma^L(t+1), \mathscr{I}^L(t,\tau-1)|\varsigma^F(t+1), \mathscr{I}^F(t)) = \sum_{\varsigma'} P(\varsigma^L(t+1), \mathscr{I}^L(t,\tau)|\varsigma^F(t+1), \mathscr{I}^F(t)),$$

where $\varsigma' = \varsigma^L(t-\tau+1)$.
Note

$$P(\varsigma(t+1), \mathscr{I}^L(t,\tau)|\mathscr{I}^F(t)) = P(\varsigma^L(t+1), \mathscr{I}^L(t,\tau)|\varsigma^F(t+1), \mathscr{I}^F(t))P(\varsigma^F(t+1)|\mathscr{I}^F(t))$$

and that

$$P(\varsigma^F(t+1)|\mathscr{I}^F(t)) = \sum_{\varsigma''}\sum_{\mathscr{I}} P(\varsigma(t+1), \mathscr{I}^L(t,\tau)|\mathscr{I}^F(t)),$$

where $\varsigma'' = \varsigma^L(t+1), \mathscr{I} = \mathscr{I}^L(t,\tau)$.

Now,

$$P(\varsigma(t+1), \mathscr{I}^L(t,\tau)|\mathscr{I}^F(t)) = P(\varsigma(t+1)|\mathscr{I}^L(t,\tau), \mathscr{I}^F(t))P(\mathscr{I}^L(t,\tau)|\mathscr{I}^F(t)).$$

Then,

$$P(\varsigma(t+1)|\mathscr{I}^L(t,\tau), \mathscr{I}^F(t)) = P(z(t+1), s(t+1)|a(t), \mathscr{I}^L(t,\tau), \mathscr{I}^F(t))P(a(t)|\mathscr{I}^L(t,\tau), \mathscr{I}^F(t)).$$

Thus, we note that $P(\varsigma^L(t+1), \mathscr{I}^L(t,\tau-1)|\varsigma^F(t+1), \mathscr{I}^F(t))$ is a function of $\{P(\mathscr{I}^L(t,\tau)|\mathscr{I}^F(t))\}$, which is the result. □

Since both π^L and π^F are finite-memory policies, then both $\mathscr{I}^L(t,\tau)$ and $y^L(t)$ are t-invariant arrays of finite dimension, which enhances the potential computability of v^L. We remark that Proposition 2.2 holds for any given finite memory leader policy ρ^L and follower policy ρ^F, where ρ^F is not necessarily a response policy to ρ^L.

We now summarize how fitness measures are determined for a given finite-memory leader policy:

- Step 1: For a given leader policy, determine a *perfect memory* follower response policy that achieves the minimum expected cost for the follower, using Proposition 2.1.
- Step 2: Approximate the resulting perfect-memory follower response policy by a *finite-memory* policy.
- Step 3: Given the leader policy and the follower's approximate response policy, determine the concomitant fitness measures using Proposition 2.2.

2.3.5 Multi-Objective Genetic Algorithm

We now describe how we use a MOGA, the NSGA-II (Deb, 2002), to generate policies from which the leader will choose a most preferred policy. We have found NSGA-II

has worked well for the example presented in Section 2.4. Our framework, however, is not restricted to any specific multi-objective genetic algorithm.

Let $\{\pi^L(m), m = 1, \ldots, M\}$ be the current population of the leader's finite memory policies, and for each m, let $\pi^F(m)$ be the approximate finite memory follower's best response policy to the leader's policy $\pi^L(m)$, where M is the population size. Further, let $v_i^L(\pi^L(m), \pi^F(m))$ be the expected cost of the leader's i^{th} criterion, given the policy pair $(\pi^L(m), \pi^F(m))$.

Definition 2.1. *Policy π^L is said to dominate policy ρ^L if*

$$v_i^L(\pi^L, \pi^F) \leq v_i^L(\rho^L, \rho^F), \forall i$$

and there exists at least one i such that

$$v_i^L(\pi^L, \pi^F) < v_i^L(\rho^L, \rho^F),$$

where π^F and ρ^F are the follower's response policies to the leader's policy π^L and ρ^L, respectively.

Policy π^L is said to be non-dominated if there does not exist a policy that dominates policy π^L.

The MOGA constructs the next generation of policies from the current set of policies as follows. The MOGA encodes each policy $\pi^L(m)$ into a chromosome. In the context of our model, a natural encoding scheme is to represent the chromosome as an array of probability mass vectors. Each row corresponds to a possible finite memory history $\mathscr{I}^L(t, \tau)$, and it is a probability mass vector over the action space for $\mathscr{I}^L(t, \tau)$ that denotes the probability that action a^L is selected by the leader given finite memory history $\mathscr{I}^L(t, \tau)$. The chromosome can be viewed as a long vector by concatenating

the probability mass vector for each possible finite memory history. Since the segment of chromosome corresponding to a given finite memory history is a probability distribution, it may create issues with offspring feasibility that can be solved by the random keys method (Bean, 1994).

Random keys is a robust yet simple technique that can ensure the feasibility of offspring without disrupting the searching process, and it has been successfully used in both discrete solution spaces (Bean, 1994) and continuous solution spaces (Lin, et al., 1998 and 2004). Random keys encodes each possible solution as a vector of random numbers of length K, where K is the cardinality of $\{\cup_{\mathscr{I}^L(t,\tau)} A^L_{\mathscr{I}^L(t,\tau)}\}$. A mapping is used to decode the solution in order to calculate the fitness measure for this solution. The mapping we used normalized each segment of chromosome corresponding to a given finite memory history $\mathscr{I}^L(t,\tau)$. This encoding scheme follows the same idea as Lin, et al., (1998, 2004). As indicated in Bean (1994) and Lin, et al., (1998), the key difference between the random keys approach and simple normalization is that the unnormalized random keys are maintained and used to search for a better solution, and it will be only normalized in order to obtain its fitness measure. The MOGA determines $v^L_i, 1 \leq i \leq N$ for each chromosome, where N is the number of the leader's objectives. The tuple $(v^L_1, \ldots, v^L_N)$ serves as the fitness measure of this chromosome.

On the basis of $(v^L_1, \ldots, v^L_N)$, the population of chromosomes is partitioned into subsets called fronts, where front 1 is the set of non-dominated chromosomes, and front $k+1$ is the set of non-dominated chromosomes when the chromosomes in fronts 1 through k are removed from consideration, $k = 1, 2, \ldots$. Chromosomes in front k are given rank k. In addition, the crowding distance of each chromosome is determined within each front. The crowding distance is defined as the average Euclidean distance of a chromosome to the other chromosomes in the front, based on $(v^L_1, \ldots, v^L_N)$ as a

measure of position. Crowding distance is considered a measure of diversity for the policies, and for this measure, larger is considered better. The current generation of policies is sorted according to ranking and crowding distance.

Parents are selected from the current generation of policies, based on their ranks and crowding distances. Chromosomes with higher rank and larger crowding distance are selected to generate offspring with higher probability. The selected parents form a mating pool and generate offspring using a crossover operator. Typical crossover operators may include one-point crossover, two-point crossover, and uniform crossover. The specific selection may depend on the needs of applications, and we use one-point crossover in our numerical example for illustration purposes. For each parents pair, the crossover operator randomly exchanges a portion of the chromosomes with each other to form two new offspring. A mutation operator is also used to maintain genetic diversity from one generation to another. This operator randomly alters a certain percentage of genes in the current generation of policies. The random keys method places no restriction on the crossover operator and the mutation operator. Then the non-dominated sorting procedure is applied again on the current population and offspring population, the top M chromosomes are kept, producing the next population. The number of iterations of the algorithm is a design parameter.

2.3.6 Equilibria

We remark that there are two equilibrium conditions, one associated with each agent. With respect to the follower, assume $\bar{\pi}^F$ is the *perfect memory* response policy to a given leader policy π^L. Then, results in Proposition 2.1 imply that

$$v^F(\mathscr{I}^F(t)) = v^F(\pi^L, \bar{\pi}^F; \mathscr{I}^F(t)) \leq v^F(\pi^L, \rho^F; \mathscr{I}^F(t))$$

for all follower policies ρ^F and all $\mathscr{I}^F(t)$, where a direct application of the results of Proposition 2.2 can be used to determine $v^F(\rho^L, \rho^F; \mathscr{I}^F(t))$ for any pair of finite-memory leader-follower policies (ρ^L, ρ^F). (In reality if an optimal policy is difficult or impossible to obtain, for any given $\epsilon > 0$, Proposition 2.1 implies $v^F(\pi^L, \bar{\pi}^F; \mathscr{I}^F(t)) - \epsilon \leq v^F(\pi^L, \rho^F; \mathscr{I}^F(t))$ for all follower policies ρ^F.)

With respect to the leader, let π^F be a finite-memory approximation of the perfect memory follower's response policy $\bar{\pi}^F$ to the given leader policy π^L. Let $v^L(\pi^L, \pi^F; \mathscr{I}^L(t))$ be the vector of criterion values for the leader's multiple objectives, given (π^L, π^F) and information state $\mathscr{I}^L(t)$. Our process of determining candidate leader policies from which the leader can select the most preferred policy is intended to determine (π^L, π^F) pairs so that there exists no pair (ρ^L, ρ^F) such that

$$v_i^L(\rho^L, \rho^F; \mathscr{I}^L(t)) \leq v_i^L(\pi^L, \pi^F; \mathscr{I}^L(t)), \forall i$$

for all $\mathscr{I}^L(t)$, where ρ^L represents any leader policy and ρ^F represents a finite-memory approximation of the perfect memory follower response policy to ρ^L.

We note that by Proposition 2.1 and results in (Puterman, 1994, Chapter 6), the follower policy in the first equilibrium condition is an optimal perfect memory policy (or ϵ-optimal policy); hence, it is not possible for the follower to achieve less (or less than $\epsilon > 0$) in expected total discounted cost by deviating from this policy. With respect to the second equilibrium condition, we note that all of the follower's policies used here are finite memory approximations of the follower's optimal perfect memory response policy to the leader's policy. Further, the process of determining the leader policies does not guarantee that pairs (π^L, π^F) will be determined that satisfy the second equilibrium condition. Thus, there is no guarantee that the leader will not want to deviate from the set of resultant non-dominated leader policies. However, given a sufficient number of generations of the MOGA and a sufficiently large τ such

that the finite memory follower policy is a good approximation to an optimal perfect memory follower policy, it is reasonable to expect that good performance for the leader can be achieved by implementing the resultant set of non-dominated leader policies generated. Figure 2.1 provides an outline of the process for generating these non-dominated leader policies.

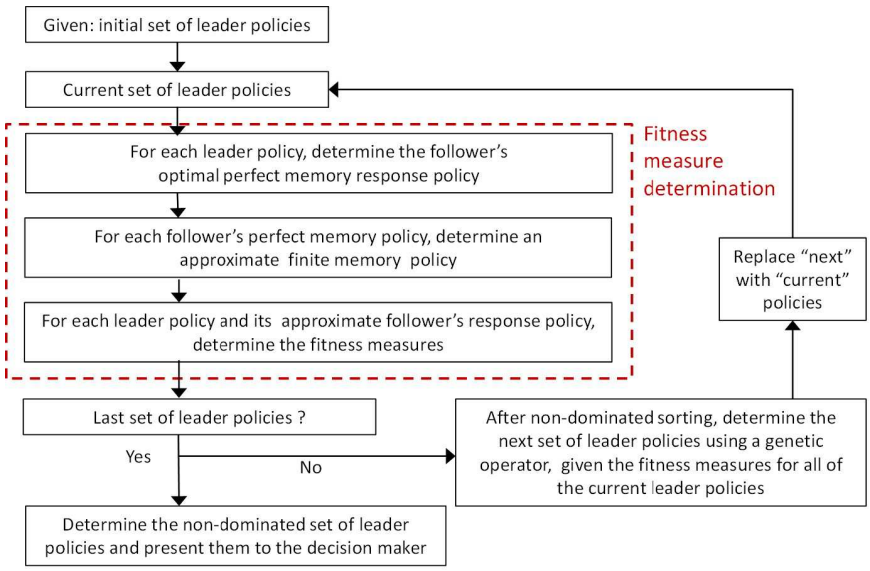


Figure 2.1: Outline of the decision support process

2.4 An Illustrative Example

We now present an example to illustrate the potential applicability and numerical feasibility of the methodology. A more in-depth application and numerical study will be presented in the future.

Liquid eggs are widely used by food service providers, such as bakers or restaurant chains, and are ingredients in many food products. We now show how the results in Section 2.3 can be used to support a production manager in selecting a sequence of actions to maximize the productivity of the liquid egg production facility while minimizing the expected number of packages that exit the facility containing a lethal

dose of a toxin due to intentional contamination. Thus, the expected number of contaminated packages represents the measure of risk under consideration.

Stackelberg games have been widely used in security applications. A software scheduling assistant called the Intelligent Randomization in Scheduling (IRIS) system was implemented in Tsai, et al., (2009) for the Federal Air Marshals (FAMs) in a single period Stackelberg game framework in order to support law enforcement aboard U.S. commercial flights. Pita, et al., (2009) developed ARMOR (Assistant for Randomized Monitoring Over Routes) using a Bayesian Stackelberg game to address the case where the defender does not know the type of adversary, and it has been successfully used at the Los Angeles International Airport to determine the checkpoints on the roadways that enter the airport and canine patrol routes within the airport terminals since 2007. However, such a single period static game cannot fully describe the dynamics of the physical environment and the adaptive nature of terrorist behavior over time, since information for each agent is updated as time proceeds and each agent will adjust its actions based upon its updated information.

The stochastic game is a powerful dynamic model for security applications, and it is especially useful when the attacker moves through successive states to "probe" a system before initiating an attack. Delle Fave, et al., (2014) developed a general Bayesian Stackelberg game model to describe the dynamic execution uncertainty for a security resource allocation problem, which can also be viewed as a special case of a stochastic Stackelberg game. Vorobeychik, et al., (2012) modelled the adversarial patrolling game as a stochastic game and computed Stackelberg equilibria. Such stochastic games assume that the agents have complete knowledge of the other agent's state, which is not always true. For example, in the food supply chain security domain, it is often the case that the defender may only observe that some restricted

toxins are missing. However, the defender may have little idea about where the toxins have gone, who is the attacker, and how the attacker may make use of the toxins. In addition, the private sector may have additional objectives (for example, maximize productivity) besides mitigating adversarial risk. Hence, we will use the framework proposed in Section 2.3 to address these issues.

We assume that both the manager and the adversary receive updated, possibly inaccurate, data about the other agent just prior to each decision epoch. The adversary's criterion is to maximize the expected number of packages that exit the facility containing a lethal dose of the toxin. Financial, morbidity, and mortality consequences would clearly be dependent on the number of contaminated liquid egg packages that exit the system, although determining these dependencies is outside of the scope of our research.

We assume that there are two targets in the production facility, $T1$ and $T2$. These two targets represent two different types of mixing tanks.

State space S^F and action space A^F of the adversary: State O is a state that the adversary occupies while assembling an attack team, manufacturing the toxin, and preparing the toxin for transport to a pre-attack state. The adversary can decide to make transition to state PT_1 and PT_2, which are the pre-attack states in which the adversary is armed and ready to attack target 1 and target 2, respectively. If in PT_i, the adversary can decide to make transition back to state O or attack target $i, i = 1, 2$. The adversary must transit through state O in order to transition from one pre-attack state to another.

State space S^L and action space A^L of the manager: The manager's states include

a full production high vulnerability state (FP), a low production low vulnerability state (LP), and the attacked state (Att). In state FP, the facility runs with a high level of productivity and the number of testing interruptions is small, and hence the cost of testing is small. In state LP, the manager interrupts the production process frequently (and expensively) to test for toxin, and as a result the productivity of the facility is low, relative to the productivity of FP. If the facility is attacked while in state LP, the probability of the attack being unsuccessful is high and the number of contaminated packages that exits the facility is small. The manager can decide at each decision epoch to keep the facility in its current state or try to switch to the other state, which then occurs with given probability. We assume that the manager can detect an attack if an attack has occurred. In this case, the manager will terminate the game; i.e., the model automatically makes transition to the attacked state (Att), which is an absorbing state (denoted by a dashed circle in Figure 2.2). The state of the system is the cross product of the adversary's state and the manager's state and is shown in Figure 2.2.

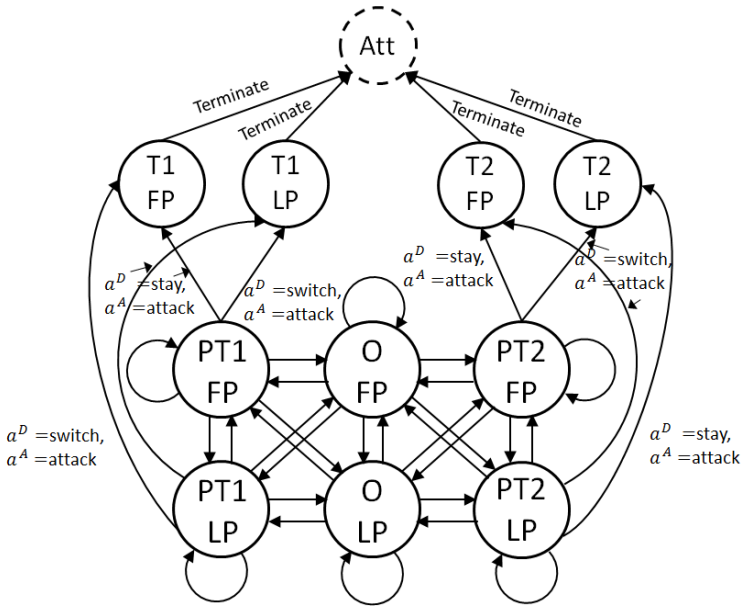


Figure 2.2: Transition diagram for the POMG numerical example

Observation spaces: There are two possible realizations of z^F, the adversary's observation of the manager: the system is well guarded; the system is not guarded. The three possible realizations of z^L, the manager's observation of the adversary, are: target 1 is under threat; target 2 is under threat; no threat. The observation matrices for the adversary and the manager are listed in Table 2.1 and Table 2.2, respectively.

Dynamics: We assume that the dynamic structure $P(s(t+1), z(t+1)|s(t), a(t))$ is given and the dynamics of the system depend on the actions of both agents. Note that $P(s(t+1), z(t+1)|s(t), a(t)) = P(z^L(t+1)|s^F(t+1))P(z^F(t+1)|s^L(t+1))P(s^F(t+1)|s^F(t), a^F(t))P(s^L(t+1)|s^L(t), a^L(t))$. Transition structures for the adversary ($P(s^F(t+1)|s^F(t), a^F(t))$) and the manager ($P(s^L(t+1)|s^L(t), a^L(t))$) can be found in Table 2.3 and Table 2.4, respectively.

Table 2.1: Adversary's observation about manager's state $P(z^F(t)|s^L(t))$

Defender's state	well guarded	not guarded	attacked
FP	0.3	0.7	0.0
LP	0.8	0.2	0.0
Att	0.0	0.0	1.0

Table 2.2: Manager's observation about adversary's state $P(z^L(t)|s^F(t))$

Attacker's state	target 1 under threat	target 2 under threat	no threat	target 1 attacked	target 2 attacked
PT_1	0.6	0.2	0.2	0.0	0.0
PT_2	0.1	0.7	0.2	0.0	0.0
O	0.4	0.4	0.2	0.0	0.0
T_1	0.0	0.0	0.0	1.0	0.0
T_2	0.0	0.0	0.0	0.0	1.0

Table 2.3: Transition structure for the adversary $P(s^F(t+1)|s^F(t), a^F(t))$

$s^F = PT1$	$PT1$	$PT2$	O	T_1	T_2
a^F="attack"	0.0	0.0	0.0	1.0	0.0
a^F="go to state O"	0.2	0.0	0.8	0.0	0.0

$s^F = PT2$	$PT1$	$PT2$	O	T_1	T_2
a^F="attack"	0.0	0.0	0.0	0.0	1.0
a^F="go to state O"	0.0	0.2	0.8	0.0	0.0

$s^F = O$	$PT1$	$PT2$	O	T_1	T_2
a^F="go to State $PT1$"	0.8	0.0	0.2	0.0	0.0
a^F="go to State $PT2$"	0.0	0.9	0.1	0.0	0.0

Table 2.4: Transition structure for the manager $P(s^L(t+1)|s^L(t), a^L(t))$

$s^L = FP$	FP	LP	Att
a^L="stay"	1.0	0.0	0.0
a^L="go to LP"	0.5	0.5	0.0
a^L="terminate and clean"	0.0	0.0	1.0

$s^L = LP$	FP	LP	Att
a^L="stay"	0.0	1.0	0.0
a^L="go to FP"	1.0	0.0	0.0
a^L="terminate and clean"	0.0	0.0	1.0

Reward structure: The reward structure for the manager has two components: the single period productivity measure r_1^L and the risk measure r_2^L, which is the number of units of contaminated liquid eggs products generated by the system if an attack

has occurred and it is zero otherwise. The reward for the adversary is $r^F(s(t), a(t)) = -\rho r_2^L(s(t), a(t))$, where ρ is a coefficient to indicate that the value of an attack for the adversary may be different from the value for the manager. The reward structure $r^k(s(t), a(t)), k \in \{L, F\}$ is due to Zhang (2013) who simulated the liquid eggs production process by a state-space model. No reward is generated once the system is in the trapping state and the game is terminated.

Information pattern and policy for the manager: The information pattern for the manager is $\mathscr{I}^L(t, \tau) = \{z^L(t), s^L(t)\}$. Thus, π^L is of the form $\{P(a^L(t)|z^L(t), s^L(t))\}$.

The construction of the POMDP for the adversary, given a manager's policy π^L: The information pattern for the adversary is $\mathscr{I}^F(t) = \{z^F(t), \ldots, z^F(1), s^F(t), \ldots, s^F(0), a^F(t-1), \ldots, a^F(0), y^F(0)\}$ when $t \geq 1$, where $y^F(0) = \{P(z^L(0), s^L(0))\}$. v^F uniquely satisfies

$$v^F(\mathscr{I}^F(t)) = \max_{a^F(t)} E\{r^F(s(t), a(t)) + \beta v^F(\mathscr{I}^F(t+1))|\mathscr{I}^F(t), a^F(t)\} \tag{2.2}$$

According to Proposition 2.1, there is an at most countable set of arrays $\Gamma^*(s^F(t))$ that only depends on $s^F(t)$, such that:

$$v^F(s^F(t), y^F(t)) = \max\{\sum_{z^L(t), s^L(t)} \gamma(z^L(t), s^L(t)) P(\{z^L(t), s^L(t)\}|\mathscr{I}^F(t)) : \gamma \in \Gamma^*(s^F(t))\}$$

where $y^F(t) = P(\{z^L(t), s^L(t)\}|\mathscr{I}^F(t))$. The POMDP solver was implemented using the technique developed in Lin et al. (1998, 2004).

Finite-memory approximation: The resulting adversary's policy is in the form of $\pi^F : \{s^F(t), y^F(t)\} \to A^F$. We approximate the perfect memory policy with a finite-memory policy when $\tau = 1$. $y^F(t)$ can be obtained by $\{z^F(t), s^F(t), a^F(t-1), y^F(t-1)\}$ using Bayes' rule. We drew 500 samples of $y^F(t-1)$ from a uniform distribution over

$Z^L \times S^L$. For each sample, we obtained the perfect memory policy. The finite-memory policy is approximated by the sample average of these perfect memory polices. The resulting policy is of the form $P(a^F(t)|\mathscr{I}^F(t,\tau=1))$ where $\mathscr{I}^F(t,\tau=1) = \{z^F(t), s^F(t), a^F(t-1)\}$.

Fitness measure determination: Given that the manager's policy is of the form $\{P(a^L(t)|z^L(t), s^L(t))\}$ and the adversary's policy is of the form $\{P(a^F(t)|\mathscr{I}^F(t,\tau=1))\}$, the manager's value function can be evaluated by Proposition 2.2 where $g_i^*, i \in \{1 = \text{productivity measure}, 2 = \text{risk measure}\}$ satisfies:

$$
\begin{aligned}
&g_i^*(\{z^L(t), s^L(t)\}, \mathscr{I}^F(t,\tau=1)) \\
&= \sum_{a(t)} P(a^F(t)|\mathscr{I}^F(t,\tau=1)) P(a^L(t)|z^L(t), s^L(t)) \{r_i^L(s(t), a(t)) \\
&+ \beta \sum_{z(t+1), s(t+1)} g_i^*[\{z^L(t+1), s^L(t+1)\}, \mathscr{I}^F(t+1,\tau=1)] P(z(t+1), s(t+1)|s(t), a(t))\},
\end{aligned}
$$

We determine g_i^* by solving a system of linear equations. Hence, the fitness measures can by computed by $v_i^L(\mathscr{I}^L(0)) = \sum_{\mathscr{I}^F(0,\tau)} g_i^*(\{z^L(0), s^L(0)\}, \mathscr{I}^F(0,\tau)) P(\mathscr{I}^F(0,\tau)|\mathscr{I}^L(0))$ where $P(\mathscr{I}^F(0,\tau)|\mathscr{I}^L(0))$ is given.

MOGA: The chromosome in MOGA encodes a manager's policy by an array of probability mass vectors. Each row is a probability mass vector over the possible manager's actions for a possible value of $\{z^L, s^L\}$. We restrict our attention to deterministic manager policies since they are easy to implement for end-users and easy to explain. The advantages of a deterministic policy are also discussed in Paruchuri, et al., (2004) and Basilico, et al., (2009). Hence, there are only 64 manager policies in total under consideration. The MOGA used has a population size of 6, one point crossover probability 1/3, and mutation rate 5%. The MOGA can probabilistically identify the Pareto efficient policies within 5 generations. Parallel programming can be used to evaluate the fitness measures in parallel for a population of leader's policies. The

runtime results on a computer with 3.10GHz CPU are summarized in Table 2.5.

Table 2.5: Runtime results

Task	Average Time (s)
Solving a POMDP	0.3266
Finite memory policy approximation	0.0951
Fitness measure determination	0.0173
Total time for MOGA to converge	6.4731

The non-dominated set of policies $(\pi_1, ..., \pi_5)$ and two baseline policies (b_1, b_2) are listed in Table 2.6, and their corresponding performance measures (after normalization) are summarized in Table 2.7. We note that both baseline policies are dominated. Policy π_5 is the 'least risk, least productivity' policy, where the manager always selects state LP. The other non-dominated policies $(\pi_1, ..., \pi_4)$ consider various tradeoffs between risk and productivity. The manager can then decide on the most preferred policy from the non-dominated policies, based on his/her own (or the corporation's) relative importance of maximizing the productivity measure and minimizing the risk measure.

Baseline policy b_1 assumes that the manager always selects state FP, presumably with the intent of maximizing productivity. However, always selecting FP also maximizes the facility's vulnerability to an attack, which in turn reduces the total long run productivity of the facility. Baseline policy b_2 alternates between FP and LP. Interestingly, b_2 has a higher measure of productivity and a lower measure of risk than b_1 (i.e., b_2 dominates b_1), pointing out the importance of considering risk when striving for increased productivity.

Table 2.6: Non-dominated policies v.s. Baseline policies

$\{z^L(t), s^L(t)\}$	π_1	π_2	π_3	π_4
(FP, T_1 under threat)	"Go to LP"	"Go to LP"	"Stay"	"Stay"
(FP, T_2 under threat)	"Stay"	"Go to LP"	"Go to LP"	"Go to LP"
(FP, No threat)	"Go to LP"	"Go to LP"	"Stay"	"Stay"
(LP, T_1 under threat)	"Go to FP"	"Go to FP"	"Stay"	"Stay"
(LP, T_2 under threat)	"Stay"	"Go to FP"	"Stay"	"Stay"
(LP, No threat)	"Go to FP"	"Stay"	"Go to FP"	"Stay"
(FP, T_1 attacked)	"Stop & clean the system"			
(FP, T_2 attacked)	"Stop & clean the system"			
(LP, T_1 attacked)	"Stop & clean the system"			
(LP, T_2 attacked)	"Stop & clean the system"			

$\{z^L(t), s^L(t)\}$	π_5	b_1	b_2
(FP, T_1 under threat)	"Go to LP"	"Stay"	"Go to LP"
(FP, T_2 under threat)	"Go to LP"	"Stay"	"Go to LP"
(FP, No threat)	"Go to LP"	"Stay"	"Go to LP"
(LP, T_1 under threat)	"Stay"	"Go to FP"	"Go to FP"
(LP, T_2 under threat)	"Stay"	"Go to FP"	"Go to FP"
(LP, No threat)	"Stay"	"Go to FP"	"Go to FP"
(FP, T_1 attacked)	"Stop & clean the system"		
(FP, T_2 attacked)	"Stop & clean the system"		
(LP, T_1 attacked)	"Stop & clean the system"		
(LP, T_2 attacked)	"Stop & clean the system"		

Table 2.7: Decision Support Table

Policy	Productivity ratio to maximum	Vulnerability ratio to minimum
π_1	1.000	7.077
π_2	0.959	5.737
π_3	0.942	5.633
π_4	0.847	4.136
π_5	0.691	1.000
b_1	0.448	45.461
b_2	0.686	9.274

2.5 Conclusions

We have blended the POMG, the POMDP, and the MOGA to identify leader policies that are candidates for a most preferred policy in an infinite horizon, partially observed, multi-period decision making environment where:

- There are two intelligent and adaptable agents, a leader and a follower, and each can affect the performance of the other.
- Before the game begins, the leader considers multiple objectives in selecting its policy. The follower knows the leader's policy and determines a response policy that is optimal with respect to the follower's objective. The policies are fixed throughout the game once selected.
- At each decision epoch, each agent knows: its past and present states, its past actions, and the possibly inaccurate observations of the other agent's past and present states.
- Each agent's policy selects actions based on data currently available to the

agent, and the actions are selected simultaneously at each decision epoch.

We have extended Stackelberg equilibrium and multi-objective decision making to the POMG. Assuming that the follower selects its policy with complete knowledge of and in response to the policy selected by the leader, we have constructed a specially structured POMDP that leads to the determination of a perfect-memory optimal policy for the follower (Proposition 2.1). We have shown that this POMDP has a computationally useful sufficient statistic and a value function structure described in terms of this sufficient statistic. By assuming that the leader policy is a finite-memory policy, we have shown that the sufficient statistic is finite-dimensional and that at least a near-optimal perfect-memory policy for the follower is potentially computable. Using a simple ad-hoc procedure for finding a finite-memory approximation to a perfect-memory policy, we have found a finite-memory policy for the follower. Assuming the policies for both agents are finite-memory policies, we have determined a computationally tractable procedure for calculating the fitness measures for the MOGA (Proposition 2.2). We then are able to compute the fitness measures for each current generation leader policy. The MOGA uses these fitness measures to create the next generation of leader policies. The non-dominated set of the final generation of leader policies created by the MOGA and the fitness measures of each non-dominated leader policy then represents the output of our process.

The output of the process described in this chapter can serve as the options generation phase of an option selection process; e.g., a deterministic version of multi-attribute decision theory (Kenney and Raiffa, 1993). Topics for future research include real-world applications of the model, an in-depth numerical analysis, and a study of the value of information, which we will introduce in next chapters.

2.6 References

[1] Aberdeen, D. A. (2003) (revised) survey of approximate methods for solving partially observable Markov decision processes, *Technical report*, Research School of Information Science and Engineering, Australia National University.

[2] Bakir, N. O. (2011) A Stackelberg game model for resource allocation in cargo container security, *Annals of Operations Research*, 187: 5 - 22.

[3] Bakir, N. O. and Kardes, K. (2009) A Stochastic game model on overseas cargo container security, *Non-published Research Reports*, CREATE center, Paper 6.

[4] Basilico N., Gatti, N., and Amigoni F. (2009) Developing a deterministic patrolling strategy for security agents, in *Proceedings of the IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT)*, 565-572.

[5] Bean, J. C. (1994) Genetic algorithms and random keys for sequencing and optimization, *ORSA Journal on Computing*, 6: 154 - 160.

[6] Becker, R., Zilberstein, S., Lesser, V., and Goldman, C. V. (2004) Solving transition independent decentralized Markov decision processes, *Journal of Artificial Intelligence Research (JAIR)*, 22: 423 - 455.

[7] Bernstein, D. S., Givan, R., Immerman, N., and Zilberstein, S. (2002) The complexity of decentralized control of Markov decision processes, *Mathematics of Operations Research*, 27(4): 819 - 840.

[8] Bernstein, D. S., Hansen, E. A., and Zilberstein, S. (2005) Bounded policy iteration for decentralized POMDPs, In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI)*, 1287 - 1292, Edinburgh, Scotland.

[9] Bier, V. M., Oliveros, S., and Samuelson, L. (2007) Choosing what to protect: Strategic defensive allocation against an unknown attacker, *Journal of Public Economic Theory*, 9(4): 563 - 587.

[10] Bier, V. M., Haphuriwat, N., Menoyo, J., Zimmerman, R., and Culpen, A. M. (2008) Optimal resource allocation for defense of targets based on differing measures of attractiveness, *Risk Analysis*, 28(3): 763 - 770.

[11] Bopardikar, S. D., and Hespanha, J. P. (2011) Randomized solutions to partial information dynamic zero-sum games, In *American Control Conference (ACC)*, San Francisco, CA.

[12] Bowman, M., Briand, L. C., and Labiche, Y. (2010) Solving the class responsibility assignment problem in object-oriented analysis with multi-objective genetic algorithms, *IEEE Transactions on Software Engineering (TSE)*, 36(6): 817 - 837.

[13] Cardoso,J.M.P, Diniz, P.C. (2009) Game Theory Models of Intelligent Actors in Reliability Analysis: An Overview of the State of the Art, *Game Theoretic Risk Analysis of Security Threats, International Series in Operations Research & Management Science*, 128: 1-19.

[14] Canu, A. and Mouaddib, A. I. (2011) Collective decision-theoretic planning for planet exploration, In *Proceedings of International Conference on Tools with Artificial Intelligence*.

[15] Cassandra, A. R., Kaelbling, L. P., and Littman, M. L. (1994) Acting optimally in partially observable stochastic domains , in *Proceedings Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, 1023 - 1028.

[16] Cassandra, A. R. (1994) Optimal policies for partially observable Markov decision processes, *Technical Report (CS-94-14)*, Brown University, Department of Computer Science, Providence RI.

[17] Cassandra, A. R., Littman, M. L. and Zhang, N. L. (1997) Incremental pruning: a simple, fast, exact method for partially observable Markov decision processes, in *Proceedings Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-97)*, Morgan Kaufmann, San Francisco, CA, 54 - 61.

[18] Cavusoglu, H., and Kwark, Y. (2013) Passenger profiling and screening for aviation security in the presence of strategic attackers, *Decision Analysis*, 10(1): 63 - 81.

[19] Cheng, H. T. (1988) Algorithms for partially observable Markov decision processes, *PhD thesis*, University of British Columbia, Vancouver, British Columbia, Canada.

[20] Coello, C. A. C. (2000) An updated survey of GA-Based multiobjective optimization techniques, *ACM Computing Survey*, 32(2): 109 - 143.

[21] Corne, D. W., Knowles, J. D. and Oates, M. J. (2000) The Pareto envelope-based selection algorithm for Multiobjective optimization, In *Proceedings of the Parallel Problem Solving from Nature VI Conference*, 1917: 839 - 848.

[22] Deb, K. (2001) Nonlinear goal programming using multi-objective genetic algorithms, *Journal of the Operational Research Society*, 52(3), 291 - 302.

[23] Deb, K., Pratap, A., Agarwal, S., and Meyarivan, T. (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II, *IEEE Transactions on Evolutionary Computation*, 6(2): 182 - 197.

[24] Delle Fave F. M., Jiang, A. X., Yin, Z. Zhang, C., Tambe, M., Kraus, S., Sullivan, J. P. (2014) Game-theoretic security patrolling with dynamic execution uncertainty and a case study on a real transit system, *Journal of Artificial Intelligence Research*, to appear.

[25] Doshi, P. (2012) Decision making in complex multiagent contexts: a tale of two frameworks, *AI Magazine*, 33(4): 82 - 95.

[26] Eagle, J. N. (1984) The optimal search for a moving target when the search path is constrained, *Operations Research*, 32(5): 1107 - 1115.

[27] Emery-Montemerlo, R., Gordon, G., Schneider, J., and Thrun, S. (2004) Approximate solutions for partially observable stochastic games with common payoffs, In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 136 - 143.

[28] Feng, Z., and Zilberstein, S. (2004) Region-based incremental pruning for POMDPs, In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence (UAI-04)* , Morgan Kaufmann, San Francisco, 146 - 153.

[29] Filar, J., and Vrieze, K. (1997) Competitive Markov decision processes, Springer, Heidelberg.

[30] Forrest, S. (1993) Genetic algorithms: principles of natural selection applied to computation, *Science*, 261: 872 - 878.

[31] Ghosh, M. K., McDonald, D., and Sinha, S. (2004) Zero-sum stochastic games with partial information, *Journal of Optimization Theory and Applications*, 121(1): 99 - 118.

[32] Gmytrasiewicz, P. J., and Doshi, P. (2005) A framework for sequential planning in multi-agent settings, *Journal of Artificial Intelligence Research*, 24: 49 - 79.

[33] Goldberg, D. E. (1989) Genetic algorithms in search, optimization, and machine learning, Addison-Wesley: Reading, MA.

[34] Hansen, E. A. (1998a) An improved policy iteration algorithm for partially observable MDPs, *Advances in Neural Inform. Processing Systems (NIPS-97)*, 10: 1015 - 1021, MIT Press, Cambridge, MA.

[35] Hansen, E. A. (1998b) Solving POMDPs by searching in policy space, in *Proceedings of Uncertainty in Artificial Intelligence*, 10: 211 - 219.

[36] Hansen, E. A., Bernstein, D. S., and Zilberstein, S. (2004) Dynamic programming for partially observable stochastic games, in *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, 709 - 715, San Jose, California.

[37] Hausken, K., and Zhuang, J. (2011) Governments' and terrorists' defense and attack in a T-period game, *Decision Analysis*, 8(1): 46 - 70.

[38] Hauskrecht M. (1997) Planning and control in stochastic domains with imperfect information, PhD thesis, Massachusetts Institute of Technology.

[39] Hauskrecht M. (2000) Value-function approximations for partially observable Markov decision processes, *Journal of Artificial Intelligence Research*, 13: 33 - 94.

[40] Hespanha, J. P., and Prandini, M. (2001) Nash equilibria in partial-information games on Markov chains, in *IEEE Conference on Decision and Control*, Orlando, FL, 2102 - 2107.

[41] Holland, J. H. (1975) Adaptation in natural and artificial systems, University of Michigan Press: Ann Arbor,. Reprinted in 1992 by MIT Press, Cambridge MA.

[42] Holloway, H, and White, C. C. (2008) Question Selection and Resolvability for Imprecise Multi-Attribute Alternative Selection, *IEEE Transactions on Systems, Man, and Cybernetics*, Part A, 38(1): 162-169.

[43] Horn, J., Nafpliotis, N., and Goldberg, D. E. (1994) A niched Pareto genetic algorithm for multiobjective optimization, In *Proceedings of the 1st IEEE conference on evolutionary computation, IEEE World Congress on Computational Intelligence* 1: 82 - 87, Orlando, FL.

[44] Kandori, M. and Obara, I. (2010) Towards a belief-based theory of repeated games with private monitoring: an application of POMDP, manuscript.

[45] Keeney, R. L. and Raiffa, H. (1993) Decisions with multiple objectives: preferences and value trade-offs, Cambridge University Press.

[46] Konak, A., Coit, D. W., and Smith, A. E. (2006) Multi-objective optimization using genetic algorithms: A tutorial, *Reliability Engineering and System Safety*, 91: 992 - 1007.

[47] Kumar, A., and Zilberstein, S. (2009) Dynamic programming approximations for partially observable stochastic games, In *Proceedings of the Twenty-Second International FLAIRS Conference*, 547 - 552, Sanibel Island, Florida.

[48] Letchford, J., Macdermed, L., Conitzer, V., Parr, R., and Isbell, C. L. (2012) Computing Stackelberg strategies in stochastic games, *ACM SIGecom Exchanges*, 11(2): 36 - 40.

[49] Lin, A. Z.-Z.,Bean, J., and White, C. C. (1998) Genetic algorithm heuristics for finite horizon partially observed Markov decision problems, *Technical Report*, University of Michigan, Ann Arbor.

[50] Lin, A. Z.-Z.,Bean, J., and White, C. C. (2004) A hybrid genetic/optimization algorithm for finite horizon partially observed Markov decision processes, *Journal on Computing*, 16(1): 27 - 38.

[51] Lin, C. M. and Gen, M. (2008) Multi-criteria human resource allocation for solving multistage combinatorial optimization problems using multiobjective hybrid genetic algorithm, *Expert Systems with Applications*, 34(4): 2480 - 2490.

[52] Littman, M. L. (1994a) The Witness algorithm: solving partially observable Markov decision processes, Brown University, Department of Computer Science, *Technical Report* CS-94-40.

[53] Littman, M. L. (1994b) Memoryless policies: theoretical limitations and practical results, In *Proceedings of the Third International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, 238 - 245.

[54] Lovejoy, W. S. (1991) A survey of algorithmic methods for partially observed Markov decision process, *Annals of Operations Research*, 28(1): 47 - 65.

[55] Manning, L., Baines, R., and Chadd, S. (2005) Deliberate contamination of the food supply chain, *British Food Journal*, 107(4): 225 - 245.

[56] McEneaney, W. M. (2004) Some classes of imperfect information finite state-space stochastic games with finite-dimensional solutions, *Applied Mathematics and Optimization*, 50(2): 87 - 118.

[57] Monahan, G. E. (1982) A survey of partially observable Markov decision processes: Theory, models, and algorithms, *Management Science*, 28(1): 1 - 16.

[58] Mohtadi, H. and Murshid, A. P. (2009) Risk analysis of chemical, biological, or radionuclear threats: Implications for food security, *Risk Analysis*, 29: 1317 - 1335.

[59] Naser-Moghadasi, M. (2012) Evaluating effects of two alternative filters for the incremental pruning algorithm on quality of POMDP exact solutions, *International Journal of Intelligence Science*, 2(1): 1 - 8.

[60] Oliehoek, F. A., Spaan, M. T. J., and Vlassis, N. (2005) Best-response play in partially observable card game, In *Proceedings of the 14th Annual Machine Learning Conference of Belgium and the Netherlands*, 45 - 50.

[61] Oliehoek, F. A., Spaan, M. T. J., and Vlassis, Nikos (2008) Optimal and approximate Q-value functions for decentralized POMDPs, *Journal of Artificial Intelligence Research*, 32: 289 - 353.

[62] Oliehoek, F. A. (2012) Decentralized POMDPs, In: M. Wiering, & M. V. Otterlo (Eds.), *Reinforcement learning: State of the art*, 12: 471 - 503, Springer.

[63] Ombuki, B., Ross, B. J., and Hanshar, F. (2006) Multi-objective genetic algorithms for vehicle routing problem with time windows, *Applied Intelligence*, 24: 17 - 30.

[64] O'Ryan, M., Djuretic, T., Wall, P., Nichols, G., Hennessy, T., Slutsker, L., Hedberg, C., MacDonald, K., and Osterholm, M. (1996) An outbreak of salmonella infection from ice cream, *New England Journal of Medicine*, 335(11): 824 - 825.

[65] Paruchuri, P., Tambe, M., Ordonez, F., and Kraus, S. (2004) Towards a formalization of teamwork with resource constraints, In *International Joint Conference on Autonomous Agents and Multiagent Systems*, 596 - 603.

[66] Pineau, J., Gordon, G. J., and Thrun, S. (2003) Point-based value iteration: An anytime algorithm for POMDPs, In *International Joint Conference on Artificial Intelligence*, 1025 - 1032.

[67] Pita, J., Jain, M., Ordonez, F., Portway, C., Tambe, M., Western, C., Paruchuri, P. and Kraus, S. (2009) Using game theory for Los Angeles airport security, *AI Magazine*, 43 - 57.

[68] Platzman, L. K. (1977) Finite memory estimation and control of finite probabilistic systems, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

[69] Platzman, L. K. (1980) Optimal infinite-horizon undiscounted control of finite probabilistic systems, *SIAM Journal on Control and Optimization*, 18: 362 - 380.

[70] Ponnambalam, S. G., Ramkumar, V., and Jawahar, N. (2001) A multiobjective genetic algorithm for job shop scheduling, *Production Planning & Control: The Management of Operations*, 12(8): 764 - 774.

[71] Poupart, P., and Boutilier, C. (2004) Bounded finite state controllers, In *Advances in Neural Information Processing Systems (NIPS) 16: Proceedings of the 2003 Conference*, MIT Press.

[72] Poupart, P. (2005) Exploiting structure to efficiently solve large scale partially observable Markov decision processes, PhD thesis, Department of Computer Science, University of Toronto.

[73] Puterman, M. L. (1994) Markov decision processes: discrete dynamic programming, New York: J Wiley & Sons.

[74] Rabinovich, Z., Goldman, C. V., and Rosenschein, J. S. (2003) The complexity of multiagent systems: the price of silence, In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 1102 - 1103, Melbourne, Australia.

[75] Raghavan, T. E. S., and Filar, J. A. (1991) Algorithms for stochastic games - a survey, *Methods and Models of Operations Research*, 35: 437 - 472.

[76] Rothschild, C., McLay, L., Guikema, S. (2012) Adversarial risk analysis with incomplete information:A level-k approach, *Risk Analysis*, 32(7): 1219 - 1231.

[77] Schaffer, J. D. (1985) Multiple objective optimisation with vector evaluated genetic algorithm, In *Proceedings of the 1st International Conference on Genetic Algorithms*, 93 - 100, Morgan Kaufmann Publishers, Inc., San Mateo.

[78] Seuken, S., and Zilberstein, S. (2007) Improved memory-bounded dynamic programming for decentralized POMDPs, In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence*, Vancouver, Canada.

[79] Shani, G., Pineau, J., and Kaplow, R. (2013) A survey of point-based POMDP solvers, *Autonomous Agents and Multi-Agent Systems*, 27:151.

[80] Shapley, L. S. (1953) Stochastic games, *Proceedings of the National Academy of Sciences of the U. S. A.*, 39: 1095 - 1100.

[81] Smallwood, R. D., and Sondik, E. J. (1973) The optimal control of partially observable Markov decision processes over a finite horizon, *Operations Research*, 21(5): 1071 - 1088.

[82] Sobel, J., Khan, A., and Swerdlow, D. (2002) Threat of a biological terrorist attack on the US food supply: the CDC perspective, *The Lancet*, 359(9309): 874 - 880.

[83] Sondik, E. J. (1978) The optimal control of partially observable Markov processes over the infinite horizon: discounted costs, *Operations Research*, 26(2): 282 - 304.

[84] Srinivas, M., and Patnaik, L. M. (1994) Genetic algorithms: A survey, *IEEE Computer*, 27(6): 17 - 26.

[85] Tsai, J., Rathi, S., Kiekintveld, C., Ordonez, F. and Tambe, M. (2009) IRIS a tool for strategic security allocation in transportation networks, *Non-published Research Report*, Paper 71, CREATE Research Archive.

[86] Ummels, M. (2010) Stochastic multiplayer games: theory and algorithms, PhD thesis, RWTH Aachen University.

[87] Vorobeychik, Y., and Singh, S. (2012) Computing Stackelberg equilibria in discounted stochastic games, In *Twenty-Sixth National Conference on Artificial Intelligence*.

[88] Vorobeychik, Y., An, B., and Tambe, M. (2012) Adversarial patrolling games, In *AAAI Spring Symposium on Security, Sustainability, and Health*.

[89] Vorobeychik, Y., An, B., Tambe, M., and Singh, S. (2014) Computing solutions in infinite-horizon discounted adversarial patrolling games, In *International Conference on Automated Planning and Scheduling*.

[90] Wang, C. and Bier, V. M. (2011) Target-hardening decisions based on uncertain multiattribute terrorist utility, *Decision Analysis*, 8(4): 286 - 302.

[91] White, C. C. (1991) A survey of solution techniques for the partially observed Markov decision process, *Annals of Operations Research*, 32(1): 215 - 230.

[92] White, C. C., and Scherer, W. T. (1989) Solution procedures for partially observed Markov decision processes, *Operations Research*, 37(5): 791 - 797.

[93] White, C. C., and Scherer, W. T. (1994) Finite-memory suboptimal design for partially observed Markov decision processes, *Operations Research*, 42(3): 439 - 455.

[94] Yildirim, M. B., and Mouzon, G. (2012) Single-machine sustainable production planning to minimize total energy consumption and total completion time using a multiple objective genetic algorithm, *IEEE Transactions on Engineering Management*, 59(4): 585-597.

[95] Yu, H. (2007) Approximation solution methods for partially observable Markov and semi-Markov decision processes, PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

[96] Zhang, H. (2010) Partially observable Markov decision processes: a geometric technique and analysis, *Operations Research*, 58(1): 214 - 228.

[97] Zhang, Y. (2013) Contributions in supply chain risk assessment and mitigation, *PhD Thesis*, Georgia Institute of Technology.

[98] Zhuang J. and Bier, V. M. (2007) Balancing terrorism and natural disasters defensive strategy with endogenous attacker effort, *Operations Research*, 55(5): 976 - 991.

CHAPTER III

VALUE OF INFORMATION FOR A LEADER-FOLLOWER PARTIALLY OBSERVED MARKOV GAME

3.1 Introduction

The research in this chapter considers the following multi-period stochastic game. There are two intelligent and adaptive decision-making agents, a leader and a follower. These agents interact as follows:

- Before the game begins, the follower chooses its policy with complete knowledge of what policy the leader has chosen.
- Once the game begins, the policies chosen by the leader and follower determine what actions are simultaneously selected at each decision epoch of an at most countable number of decision epochs.
- The decisions of both policies affect the dynamics of the system that both agents want to control.
- Each agent's policy makes decisions to achieve its agent's objective, based on data that include possibly inaccurate, incomplete, and/or costly observations of the other agent.

The question addressed in this chapter is: how does the accuracy of the leader's observation of the follower affect the performance of the leader? A related question: what is the added value to the leader if more accurate (and presumably, requiring greater resources) observations of the follower could be made available? Another related

question: is it guaranteed that more accurate information about the follower will improve leader performance? We use a leader-follower partially observed Markov game (POMG) to model and address these "value of information" questions by building on previous research presented in Chapter 2, specialized to the case where the leader has a single objective (the leader was assumed to have multi-objectives in Chapter 2.

It seems intuitive that more accurate observations of the underlying state of a system subject to control will lead to better system performance. Better medical diagnostic quality should lead to healthier patient outcomes; higher quality machine fault identification should lead to more effective machine maintenance and better system performance; more accurate observations of an adversary or a competitor should provide advantage. However, several counterexamples to this claim for the partially observed Markov decision process (POMDP), a single-agent decision making model, have been found in Ortiz, Erera and White (2013).

The assumption that the follower has complete knowledge of the leader's policy is realistic in many applications and can serve as a worst-case scenario ("erring" in the right direction) otherwise. With respect to realism, if the leader is defending a key infrastructure site and the follower's objective involves breaching site security, the follower may have at least partial knowledge of the leader's policy by observing the leader's allocation of defensive resources. Also, the follower may be able to infer, at least in part, the leader's policy from sources such as publicly available information if, for example, the leader is associated with a government agency with open records. We remark that the underlining state of each agent is only partially observed by the other agent and hence each agent's policy selects actions based on data different from the other agent's data. Thus, in general the follower will not know exactly what action the leader will take, which may mollify concern in situations where the leader-follower

assumption is believed to unrealistically bias results to the advantage of the follower.

The intent of the research is to better inform the decision to seek or not seek improved state observation quality and what resulting changes in performance to expect if state observation quality is improved. The initial motivation for this research was how to support a manager of a food supply chain (the leader) in selecting a sequence of actions that best balances maximizing the performance of, while minimizing the risk to, the supply chain over time, given that there is an attacker (the follower) who seeks to contaminate the food supply chain with a chemical or biological toxin. More specifically in the context of this application, if the manager (or a government agency) is able to improve the quality of observing the attacker (presumably at a cost), is it useful to do so? Details of this application can be found in Chapter 2. A detailed discussion of the reasons behind the interest in measures of risk and risk mitigation when an adversary is intelligent and adaptive can be found in Ezell, et al., (2010).

This chapter is organized as follows. Section 3.2 presents a review of the literature. We consider the single agent case in Section 3.3. This case and its associated value determination are important for results to follow and are the basis of the solution procedure for the POMG found Chapter 2 because:

- Given a leader policy, the (two-agent) POMG can be transformed into a (single agent, potentially tractable) POMDP that determines the follower's response policy.
- Computing the value of either the leader's objective function or the follower's objective function requires both a leader's policy and a follower's policy.

The POMDP problem statement and preliminary results from the literature are presented in Section 3.3.1 through 3.3.3. We summarize two definitions of observation

quality and related results from the literature. Theorem 3.1 and Corollary 3.1 present conditions that guarantee that improved observation quality insures improved system performance for the first definition. Results with respect to the second definition show that improved observation quality may degrade system performance. We then compare and contrast these two definitions in Section 3.3.4.

Section 3.4 considers the POMG. Two descriptions of the value functions useful for later results are presented in Section 3.4.2, following the definition of the POMG in Section 3.4.1. In Section 3.4.3, we present a partition of the set of all observation matrices that describe the leader's observations of the follower's underlying state. Each element of this partition contains observation matrices that share a common follower policy. Section 3.4.4 studies the impact on the leader's value function of changing observation quality within a partition element. We (1) extend the POMDP results to the POMG under these circumstances, (2) show that the value function is Lipschitz continuous in observation quality, and then (3) present conditions that guarantee that if an observation matrix is a member of one of the partition elements and a second observation is sufficiently close to the first observation matrix, then the second observation matrix is also a member of the same partition element. In Section 3.4.5, we examine the implications of changing the observation matrix to an observation matrix in a different partition element. We show by example that crossing partition element boundaries may produce discontinuities in the leader's value function. We give conditions that guarantee that a discontinuity will be favorable to leader performance when such a discontinuity occurs. We show that when both agents are collaborative and the follower initially has complete observability of the leader, then discontinuities cannot occur. However, we show by example that improving observation quality does not necessarily improve the leader's value function, whether or not the POMG is a collaborative game.

3.2 Literature Review

The POMDP is a sequential decision making model involving a single decision maker. Compared to the completely observed Markov decision process (i.e., the MDP; see Puterman 1994), the POMDP takes into consideration inaccurate, incomplete, and/or costly observations of the state of the system under control. However, the POMDP also introduces significant computational challenges. In the past decades, structural properties of the value function and computational procedures for the POMDP have been investigated and can be found in Smallwood and Sondik (1973), Sondik (1978), Monahan (1982), White and Scherer (1989), Lovejoy (1991), White (1991), Littman (1994), Cassandra, Kaelbling and Littman (1994), Cassandra, Littman and Zhang (1997), Lin, Bean and White (1998, 2004), and Zhang (2010). Finite memory controllers for the POMDP have also been examined by Platzman (1977, 1980), White and Scherer (1994), Meuleau et al. (1999), and Poupart and Boutilier (2004). The finite memory controller for the POMDP in White and Scherer (1994) is extended to the POMG in this chapter.

The value of information for the POMDP has been addressed in White and Harrington (1980), Zhang (2010) and Ortiz, Erera and White (2013). This research has shown that for some sub-optimal policies, the value function may not necessarily improve if provided with more accurate state observations. We remark that there are results in the decision analysis literature (e.g., Wakker, 1988) that address situations where the value of information may be negative. A comparison of the results in the POMDP literature with the condition in Wakker (1988) is a topic of future research. The definitions of observation quality presented in the POMDP literature are compared and used in this chapter in order to study the value of information for POMG.

The stochastic game (Shapley, 1953) is a dynamic game that is played in a sequence of stages. The state of the system evolves probabilistically over time and is controlled by one or more players. The partially observed stochastic game (POSG) is a new and relatively unexamined generalization of a stochastic game that represents a multi-agent sequential decision making problem, where the states of the game are not precisely observed by the agents and all agents make a sequence of decisions based on these partial observations. Not surprisingly, the increased modeling realism of the POSG has resulted in increased computational challenges (see Bernstein et al. 2002, Rabinovich, Goldman and Rosenschein 2003). A dynamic program presented in Hansen, Bernstein and Zilberstein (2004) and Kumar and Zilberstein (2009) was used to prune very weakly dominated policies for both agents. Chapter 2 developed a solution procedure that can generate a set of non-dominated policies for a partially observed multi-objective Markov game, from which one of two agents (the leader) can select a most preferred policy to control a dynamic system that is also affected by the control decisions of the other agent (the follower).

The value of information is a topic of considerable interest in the game theory and decision analysis literature. Much of this interest has focused on determining the value of information in games having a single decision epoch and has shown that the reward can be either improved or degraded when the decision makers are given more accurate information. For example, research by Li (2002), Chu and Lee (2006) and Leng and Parlar (2009) have examined the positive value of revealing private information to coordinate among players in the context of supply chain management problems. Bier, Oliveros and Samuelson (2007) also examined the advantage of revealing private information to an attacker in a homeland security context. Bassan et al. (2003) identified a class of games where the value of information is positive.

Lehrer and Rosenberg (2006) and Meyer, Lehrer and Rosenberg (2010) showed that the value of information is always positive in a two-person zero-sum game. On the contrary, Kamien, Tauman and Zamir (1990) presented an example of a card game in extensive form that has a negative value of information. Zhuang and Bier (2010) argued that it is better not to reveal private information in a homeland security resource allocation problem. In multiple period games, Lehrer and Rosenberg (2010) showed that the value of information is positive for zero-sum repeated games. Zhuang, Bier, and Alagoz (2010) illustrated the use of secrecy and deception in a multiple-period, attacker-defender signaling game. To the best of our knowledge, this chapter is the first to analyze the value of information of general-sum partially observed stochastic games. We study the value of improving state observation quality within the POMG framework presented in Chapter 2.

3.3 The Single Agent Case

3.3.1 POMDP Problem Statement

Let $\{s(t), t = 0, 1, \}, \{z(t), t = 1, 2, \}$, and $\{a(t), t = 0, 1, \}$ be the state, observation, and action processes, each having finite state spaces S, Z, and A, respectively. The conditional probability $p_{ij}(z, a) = P[z(t+1) = z, s(t+1) = j | s(t) = i, a(t) = a]$ is assumed given. Let $P(z, a)$ be the sub-stochastic matrix $\{p_{ij}(z, a)\}$. Note that $P[z(t+1) = z, s(t+1) = j | s(t) = i, a(t) = a] = P[z(t+1) = z | s(t+1) = j, s(t) = i, a(t) = a] \times P[s(t+1) = j | s(t) = i, a(t) = a]$, where $p_{ij}(a) = P[s(t+1) = j | s(t) = i, a(t) = a] = \sum_z p_{ij}(z, a)$ and $q_{ijz}(a) = P[z(t+1) = z | s(t+1) = j, s(t) = i, a(t) = a] = \frac{p_{ij}(z,a)}{\sum_z p_{ij}(z,a)}$ (assuming $\sum_z p_{ij}(z, a) \neq 0$) are referred to as the transition and observation probabilities, respectively. Throughout, we will assume $q_{ijz}(a)$ is independent of i and hence $q_{jz}(a) = P[z(t+1) = z | s(t+1) = j, a(t) = a]$. Let $Q(a) = \{q_{jz}(a)\}$, be the observation matrix, which we will use as our model of state observation quality.

We consider two cases for selecting actions, the perfect memory case and the finite memory case. In both cases, decision epochs are countable, $t = 0, 1, \ldots$. Let $x(0) = \{x_i(0)\}$, where $x_i(0) = P[s(0) = i]$, $d(t) = \{z(t), \ldots, z(1), a(t-1), \ldots, a(0)\}$, and $d(t, \tau) = \{z(t), \ldots, z(t-\tau+1), a(t-1), \ldots, a(t-\tau)\}$, where τ is fixed and finite and can be thought of as a design parameter. For the perfect memory case, $a(0)$ is selected on the basis of $x(0)$ and for $t \geq 1$, $a(t)$ is selected on the basis of $\{d(t), x(0)\}$. For the finite memory case, $a(t)$ is selected on the basis of $d(t, \tau)$, where $d(0, \tau)$ is assumed given.

Let $r(i, a)$ be the reward received at epoch t, given $s(t) = i$ and $a(t) = a$. The criterion we consider for the POMDP is the infinite horizon, expected total discounted reward $E\{\sum_t \beta^t r(s(t), a(t))|x(0)\}$, where $E\{.|x(0)\}$ is the expectation operator conditioned on $x(0)$ and where we will assume the discount factor β is such that $0 \leq \beta < 1$. The POMDP has two fundamental objectives. First, determine the value of the criterion for any given policy. Second, determine a policy (an optimal policy) that maximizes the criterion and its criterion value. Our interest in the POMDP formulation is to determine the value of the criterion for a given policy, to understand how this value changes as the observation matrix changes, and to extend these results to the POMG where appropriate.

3.3.2 Perfect Memory Case

We now assume all policies are perfect memory policies.

Value Determination. Let $v^{\pi Q}(d(t))$ be the value of the criterion, assuming $\{d(t), x(0)\}$ (for notational simplicity we delete explicit dependence on x(0)), π is the policy under

consideration, and Q is the observation matrix. Then $v^{\pi Q}(d(t))$ satisfies the equation

$$v^{\pi Q}(d(t)) = \sum_i r(i, a(t)) P[s(t) = i|d(t)] + \beta \sum_z P[z(t+1) = z|d(t), a(t)] v^{\pi Q}(z, d(t), a(t)), \quad (3.1)$$

where $a(t) = \pi(d(t))$. A simple contraction mapping argument guarantees that Equation (3.1) has a unique solution.

Let $v^{*Q}(d(t)) = \max_\pi v^{\pi Q}(d(t))$. Results in Bertsekas (1976) show that $v^{*Q}(d(t))$ satisfies the following optimality equation

$$v^{*Q}(d(t)) = \max_a \{\sum_i r(i, a) P[s(t) = i|d(t)] + \beta \sum_z P[z(t+1) = z|d(t), a] v^{*Q}(z, d(t), a)\}. \quad (3.2)$$

Further, the perfect memory policy that causes the maximum to be attained is an optimal perfect memory policy.

Sufficient Statistic. Due to computational tractability concerns, we seek a t-invariant sufficient statistic, observing that $d(t)$ gets large as t gets large. According to Bertsekas (1976), there exists an optimal policy that depends on $d(t)$ only through $x(t) = \{x_i(t)\}$, where $x_i(t) = P[s(t) = i|d(t)]$; i.e., $\{x(t), t = 0, 1, \ldots\}$ is a sufficient statistic for the optimization problem. Furthermore, v^{*Q} depends on $d(t)$ only through $x(t)$, and hence Equation (3.2) can be transformed into

$$v^{*Q}(x) = \max_a \{xr(a) + \beta \sum_z \sigma(z, x, a) v^{*Q}(\lambda(z, x, a))\}, \quad (3.3)$$

where the policy that causes the maximum to be attained is an optimal policy, $xr(a) = \sum_i x_i r(i, a)$, $\sigma(z, x, a) = \sum_i x_i \sum_j p_{ij}(a) q_{jz}(a) = xP(z, a)1$ ($y1 = \sum_i y_i$ for any vector y), and $\lambda(z, x, a) = \{\lambda_j(z, x, a)\} = \frac{xP(z,a)}{\sigma(z,x,a)}$, where $\lambda_j(z, x, a) = \frac{\sum_i x_i p_{ij}(a) q_{jz}(a)}{\sigma(z,x,a)}$ when $\sigma(z, x, a) \neq 0$. We remark that $\sigma(z, x, a) = P[z(t+1) = z|d(t), a(t) = a]$ when $x(t) =$

x and $x(t+1) = \lambda(z, x, a)$ when $z(t+1) = z, x(t) = x$, and $a(t) = a$. Hence, $\lambda(z, x, a)$ is a form of Bayes' Rule. A result that is often exploited computationally is that the successive approximations operation preserves concavity and piecewise linearity; i.e., if v is concave and piecewise linear, then $\max_a\{xr(a) + \beta \sum_z \sigma(z, x, a)v(\lambda(z, x, a))\}$ is also concave and piecewise linear. In the limit, concavity is preserved, and hence it is also true that $v^{*Q}(x)$ is concave.

If we restrict our attention to the class of perfect memory policies that depend on $d(t)$ only through $x(t)$, then it is straightforward to show that $v^{\pi Q}(d(t))$ depends on $d(t)$ only through $x(t)$ and satisfies the equation

$$v^{\pi Q}(x) = xr(\pi(x)) + \beta \sum_z \sigma(z, x, \pi(x))v^{\pi Q}(\lambda(z, x, \pi(x))). \tag{3.4}$$

State Observation Quality. We now consider the first definition of observation quality presented in White and Harrington (1980) and Zhang (2010).

Definition 3.1. *Observation matrix Q' is at least as informative as observation matrix Q if there exists a stochastic matrix R such that $Q'(a)R(a) = Q(a)$ for all a.*

Theorem 3.1. *Assume that observation matrix Q' is at least as informative (in terms of Definition 1) as observation matrix Q. Let $v^{\pi Q}(x)$ be the solution of Equation 3.4, and assume $v^{\pi Q}(x)$ is concave in x. Then, $v^{\pi Q}(x) \leq v^{\pi Q'}(x)$ for all x.*

Proof of Theorem 3.1 is given in White and Harrington (1980).

A policy is said to be Q-adaptive if when given a more informative observation matrix, the policy gives in return improved performance. Theorem 3.1 states that if $v^{\pi Q}$ is concave, then π is Q-adaptive. Theorem 3.1 leads to the following results.

Corollary 3.1. *(a) Assume π is optimal for observation matrix Q, π' is optimal for observation matrix Q', and that observation matrix Q' is at least as informative as observation matrix Q (in terms of Definition 1). Then, $v^{\pi Q}(x) \leq v^{\pi' Q'}(x)$ for all x. (b) Assume Q has rank 1 and $Q'' = I$, where I is the identity matrix, and let π and π'' be optimal policies for observation matrices Q and Q'', respectively. Let Q' be any observation matrix, and assume π' is an optimal policy for Q'. Then, $v^{\pi Q}(x) \leq v^{\pi' Q'}(x) \leq v^{\pi'' Q''}(x)$ for all x.*

Corollary 3.1(a) results from the fact that an optimal policy always produces a concave value function. Corollary 3.1(b) notes that there is an observation matrix (or family of matrices) for which any given observation matrix is at least as informative, and there is an observation matrix (or family of matrices) that is at least as informative as any given observation matrix. The POMDP based on the former observation matrix (having rank 1) is called the completely unobserved case, and the POMDP based on the latter observation matrix (the identity matrix) is called the completely observed case (or simply, the MDP). The value functions of these special cases represent lower and upper bounds, respectively, on the value function of the general case.

3.3.3 Finite Memory Case

We now state an alternative definition of observation quality presented in Ortiz, Erera and White (2013), assuming that the underlying state of the system is close to completely observed (e.g., true for many inventory systems) and that all policies considered are finite-memory.

Definition 3.2. *Let $Q(\epsilon) = I(1-\epsilon) + P\epsilon$, P is a stochastic matrix with zeros on the diagonal, and $\epsilon \geq 0$ is small. Observation matrix $Q(\epsilon')$ is at least as informative as observation matrix $Q(\epsilon)$ if and only if $\epsilon' \leq \epsilon$.*

Results in Ortiz, Erera and White (2013) state that for finite-memory policy π,

$$v^{\pi Q}(x) = \lambda^{\tau}(d(t,\tau), x')\gamma(d(t,\tau)),$$

where $x = x(t) = \lambda^{\tau}(d(t,\tau), x')$ given $x' = x(t-\tau)$ and $\gamma(d(t,\tau))$ is polynomial in ϵ; i.e., there is a sequence of (easily computed) vectors $\{\alpha^k(d(t,\tau))\}$ such that $\gamma(d(t,\tau)) = \sum_{k=0}^{\infty} \epsilon^k \alpha^k(d(t,\tau))$. Thus, for sufficiently small $\epsilon > 0$, $\sum_{k=0}^{\infty} \epsilon^k \alpha^k(d(t,\tau))$ is well-defined and can be approximated by $\alpha^0(d(t,\tau)) + \epsilon\alpha^1(d(t,\tau))$, and hence the signs of the scalar elements of the vector $\alpha^k(d(t,\tau))$ for $k = 1$ determine whether or not $v^{\pi Q}(x)$ will increase or decrease as a function of ϵ. Not unexpectedly, if the finite-memory policy under consideration achieves the maximum for the completely observed MDP, then it is shown in Ortiz, Erera and White (2013) that the signs of all elements of the vector $\alpha^1(d(t,\tau))$ are negative; hence, this policy improves system performance if given improved observation quality.

3.3.4 A Comparison of Definitions of Observation Quality

We now show that if there exists a stochastic matrix R such that $Q(\epsilon)R = Q(\epsilon')$, then for sufficiently small ϵ and ϵ', $\epsilon' \geq \epsilon > 0$. Further, we show that for sufficiently small ϵ and ϵ', $\epsilon' \geq \epsilon > 0$, there may not exist a stochastic matrix R such that $Q(\epsilon)R = Q(\epsilon')$. Thus, for an observation matrix having the specialized form $Q(\epsilon) = I(1-\epsilon) + P\epsilon$, where $\epsilon \geq 0$ is small and P is a stochastic matrix with zeros on the diagonal, the definition "$Q(\epsilon)$ is at least as informative as $Q(\epsilon')$ when $\epsilon' \geq \epsilon > 0$" is more general than the definition "$Q(\epsilon)$ is at least as informative as $Q(\epsilon')$ when there exists a stochastic matrix R such that $Q(\epsilon)R = Q(\epsilon')$".

We now determine that the existence of a stochastic matrix R such that if $Q(\epsilon)R = Q(\epsilon')$ implies $\epsilon' \geq \epsilon > 0$ for sufficiently small ϵ and ϵ'. Equivalently, if $\epsilon' < \epsilon$, then

$Q(\epsilon)^{-1}Q(\epsilon')$ cannot be stochastic, assuming the existence of $Q(\epsilon)^{-1}$. Assume $\epsilon' < \epsilon$, and let $\kappa = \frac{\epsilon}{1-\epsilon}$ and $\kappa' = \frac{\epsilon'}{1-\epsilon'}$. Then, $Q(\epsilon)^{-1}Q(\epsilon') = \frac{(1-\epsilon')}{1-\epsilon}(I+\kappa P)^{-1}(I+\kappa' P)$. Since $\kappa < 1$ for ϵ sufficiently small and $(I+\kappa P)^{-1} = I - \kappa P(I+\kappa P)^{-1}$, it follows that $(I+\kappa P)^{-1} = \sum(-1)^n \kappa^n P^n$, where the sum is over all $n = 0, 1, 2, \ldots$. It is then straightforward to show that $(I+\kappa P)^{-1}(I+\kappa' P) = I + (\kappa' - \kappa)P(I - \kappa P + \kappa^2 P^2 - \ldots)$, which can be approximated by $I + (\kappa' - \kappa)P$ for small κ. Thus, for $i \neq j$, the $(i,j)^{th}$ term of $R = Q(\epsilon)^{-1}Q(\epsilon')$ is $r_{ij} = (1-\epsilon')(\kappa' - \kappa)p_{ij}/(1-\epsilon)$, which is negative since $\epsilon' < \epsilon$, and the result is proved.

If $\epsilon' > \epsilon > 0$ (both small), does there exist a stochastic matrix R such that $Q(\epsilon)R = Q(\epsilon')$? Such a stochastic matrix exists when the state and observation spaces have cardinality 2 (let $r_{11} = r_{22} = (1 - \epsilon - \epsilon')/(1-2\epsilon)$ and $r_{12} = r_{21} = (\epsilon' - \epsilon)/(1-2\epsilon)$, for $\epsilon < \frac{1}{2}$ and $(\epsilon' + \epsilon) < 1$). However, the existence of such a stochastic matrix R is not guaranteed for larger dimensional problems. For example, let

$$Q(\epsilon) = \begin{bmatrix} 1-\epsilon & \epsilon & 0 \\ 0 & 1-\epsilon & \epsilon \\ 0 & \epsilon & 1-\epsilon \end{bmatrix}.$$

Then,

$$Q(\epsilon)^{-1} = \frac{1}{(1-\epsilon)(1-2\epsilon)} \begin{bmatrix} 1-2\epsilon & -\epsilon(1-\epsilon) & \epsilon^2 \\ 0 & (1-\epsilon)^2 & -\epsilon(1-\epsilon) \\ 0 & -\epsilon(1-\epsilon) & (1-\epsilon)^2 \end{bmatrix},$$

where the inverse exists when $\epsilon < \frac{1}{2}$. We observe that the $(1,3)$ entry of $R = Q(\epsilon)^{-1}Q(\epsilon')$ is $r_{13} = -\frac{\epsilon(\epsilon'-\epsilon)}{(1-\epsilon)(1-2\epsilon)} < 0$, and hence R, although unique, is not stochastic. Thus, if $\epsilon' > \epsilon > 0$, it is not guaranteed that there exists a stochastic matrix R such that $Q(\epsilon)R = Q(\epsilon')$.

3.4 Partially Observed Markov Game

3.4.1 Problem Statement

Thus far, we have explored the impact of changes to the observation quality of the underlying state process on system performance for the POMDP and have presented results that address the question: will improved observation quality improve system performance? Given this context, we now investigate this question for the infinite horizon, expected total discounted POMG.

We assume that the POMG has two agents. The first agent, the leader, chooses its policy. Then the second agent, the follower, selects its policy with complete knowledge of the policy selected by the leader. Let $\{s^k(t), t = 0, 1, \ldots\}$, $\{z^k(t), t = 1, \ldots\}$, and $\{a^k(t), t = 0, 1, \ldots\}$ be the state, observation, and action processes for agent $k \in \{L =$ Leader, $F =$ Follower$\}$, each having finite state spaces S^k, Z^k, and A^k, respectively. Let $s(t) = \{s^L(t), s^F(t)\}$, $z(t) = \{z^L(t), z^F(t)\}$, and $a(t) = \{a^L(t), a^F(t)\}$, where $z^k(t)$ is the observation received by agent k of the other agent's state. The conditional probability $p_{ij}(z, a) = P[z(t+1) = z, s(t+1) = j | s(t) = i, a(t) = a]$ is assumed given. Let $P(z, a)$ be the sub-stochastic matrix $\{p_{ij}(z, a)\}$.

Let the information pattern at time t of finite length τ for agent k be $d^k(t, \tau) = \{s^k(t), ..., s^k(t-\tau+1), z^k(t), ..., z^k(t-\tau+1), a^k(t-1), ..., a^k(t-\tau)\}$, hence, $d^k(t, \tau) = \{s^k(t), z^k(t), a^k(t-1), d^k(t-1, \tau-1)\}$. And let $y^k(t) = \{P(d^l(t, \tau) | d^k(t))\}$ for $l \neq k$, where $y^k(t)$ is a "belief" array that indicates what agent k can infer about the other agent's information pattern, i.e., $d^l(t, \tau), l \neq k, l, k \in \{L, F\}$. Denote $d^k(t) = \{z^k(t), ..., z^k(1), s^k(t), ..., s^k(0), a^k(t-1), ..., a^k(0), y^k(0)\}$ when $t \geq 1$, where $y^k(0) = \{P(d^l(0, \tau))\}$, hence, $d^k(t) = \{z^k(t), s^k(t), a^k(t-1), d^k(t-1)\}$. The decision epochs are $t = 0, 1, \ldots$, and agent k selects $a^k(t)$ on the basis of information pattern

$d^k(t,\tau)$. Hence, when selecting an action, we assume that agent k knows the current and τ most recent observations of the other agent's state, its current and τ most recent state values, and the τ most recent actions it has selected. Let $v^k(\pi^L,\pi^F,Q)(d^k(0))$ be the value of agent k's criterion, assuming the leader and follower policies are π^L and π^F, respectively, and $Q=\{P(z^L|s^F)\}$ is the leader's observation matrix. We let $d(t,\tau)=\{d^L(t,\tau),d^F(t,\tau)\}$. It will be convenient to describe the policy pair (π^L,π^F) as $\{P(a(t)|d(t,\tau))\}$, where $P(a(t)|d(t,\tau))=P(a^L(t)|d^L(t,\tau))P(a^F(t)|d^F(t,\tau))$.

The criterion we consider for agent k is the infinite horizon, expected total discounted reward; i.e., $v^k(\pi^L,\pi^F,Q)(d^k(0))=E\{\sum_t \beta^t r^k(s(t),a(t))|d^k(0)\}$, where $E\{.|d^k(0)\}$ is the expectation operator conditioned on $d^k(0)$, $\beta\in[0,1)$ is the discount factor, and $r^k(i,a)$ is the scalar reward received by agent k at epoch t, given $s(t)=i$ and $a(t)=a$.

Let $\mathscr{Q}$ be the set of all stochastic matrices and hence the set of all observation matrices $Q=\{P(z^L|s^F)\}$. We remark that $\mathscr{Q}$ is equivalent to the set of all elements in $R^{|S^F|}\times R^{(|Z^L|-1)}$ such that $q_{jz}\geq 0$ for all $j\in S^F$ and $z\in Z^L$, and $\sum_{z=1}^{|z^L|-1} q_{jz}\leq 1$ for all $j\in S^F$. For example, if $|S^F|=|Z^L|=2$, then $\mathscr{Q}$ is equivalent to $\{(x,y):0\leq x\leq 1,0\leq y\leq 1\}$. Hence, $\mathscr{Q}$ is compact.

Assume initial conditions $d^L(0)$ and $d^F(0)$ are given. Let Π^L and Π^F be the set of all policies from which the leader and the follower can choose, respectively. Both Π^L and Π^F are assumed to contain only finite-memory policies. Let the response function of the follower $\pi^*:\Pi^L\times\mathscr{Q}\rightarrow\Pi^F$ be such that $\forall\pi^L\in\Pi^L$,

$$v^F(\pi^L,\pi^*(\pi^L,Q))(d^F(0))\geq v^F(\pi^L,\rho^F)(d^F(0)),\forall\rho^F\in\Pi^F.$$

Then, the equilibrium conditions for the POMG are:

(i) $v^L(\pi^L,\pi^*(\pi^L,Q))(d^L(0))\geq v^L(\rho^L,\pi^*(\rho^L,Q)))(d^L(0)),\forall\rho^L\in\Pi^L$;

(ii) $\forall \rho^L \in \Pi^L, v^F(\rho^L, \pi^*(\rho^L, Q))(d^F(0)) \geq v^F(\rho^L, \rho^F)(d^F(0)), \forall \rho^F \in \Pi^F$.

Hence, neither the leader nor the follower can improve its performance by deviating from the equilibrium condition.

Our focus in this chapter is to determine $v^L(\rho^L, \pi^*(\rho^L, Q), Q)(d^L(0))$ for any given $\rho^L \in \Pi^L$. We assume that determining this scalar value is a critical step in determining the most preferred leader policy in Π^L. We note that a genetic algorithm was used in Chapter 2 to determine the most preferred leader policy, using $v^L(\rho^L, \pi^*(\rho^L, Q), Q)(d^L(0))$ as the fitness measure.

3.4.2 Descriptions of v^k, $k \in \{L, F\}$

We now present two descriptions of v^k in Proposition 3.1 and Proposition 3.2 that will be useful in determining results below. The first description describes v^k in a manner analogous to the description of the optimality equation for the POMDP. The second description takes advantage of the fact that both leader and follower policies are assumed to be finite memory policies.

Proposition 2.2 in Chapter 2 implies that $\{d^k(t, \tau), y^k(t)\}$ is a sufficient statistic for $\{d^k(t)\}$, and hence $v^k(\pi^L, \pi^F, Q)(d^k(t)) = v^k(\pi^L, \pi^F, Q)(d^k(t, \tau), y^k(t))$.

Let one-period information for agent k be $\varsigma^k(t) = \{z^k(t), s^k(t), a^k(t-1)\}$ and $\varsigma(t) = \{\varsigma^L(t), \varsigma^F(t)\}$. Define

(i) $\sigma^k(\varsigma^k(t+1), d^k(t, \tau), y^k(t)) = P(\varsigma^k(t+1)|d^k(t))$
$= \sum_{\varsigma^l(t+1)} \sum_{d^l(t,\tau)} P(z(t+1), s(t+1)|s(t), a(t)) P(a(t)|d(t, \tau)) P(d^l(t, \tau)|d^k(t, \tau)), l \neq k$

(ii) $\lambda^k(\varsigma^k(t+1), d^k(t,\tau), y^k(t))$ is the stochastic array with scalar element

$$P(\varsigma^l(t+1), d^l(t,\tau-1)|\varsigma^k(t+1), d^k(t)) = \frac{P(\varsigma(t+1), d^l(t,\tau-1)|d^k(t))}{P(\varsigma^k(t+1)|d^k(t))},$$

where

$$P(\varsigma(t+1), d^l(t,\tau-1)|d^k(t))$$
$$= \sum_{\varsigma^l(t-\tau+1)} P(z(t+1), s(t+1)|s(t), a(t))P(a(t)|d(t,\tau))P(d^l(t,\tau)|d^k(t)), l \neq k,$$

and where we assume $P(\varsigma^k(t+1)|d^k(t)) \neq 0$.

Proposition 3.1. *For policies π^L and π^F and observation matrix Q,*

$$v^k(\pi^L, \pi^F, Q)(d^k(t)) = v^k(\pi^L, \pi^F, Q)(d^k(t,\tau), y^k(t))$$
$$= \sum_{s(t)} \sum_{a(t)} r^k(s(t), a(t)) \sum_{d^l(t,\tau)} P(a(t)|d(t,\tau))P(d^l(t,\tau)|d^k(t))$$
$$+ \beta \sum_{\varsigma^k(t+1)} \sigma^k(\varsigma^k(t+1), d^k(t,\tau), y^k(t))$$
$$\times v^k(\pi^L, \pi^F, Q)(\{\varsigma^k(t+1), d^k(t,\tau-1)\}, \lambda^k(\varsigma^k(t+1), d^k(t,\tau), y^k(t))), l \neq k.$$

Proof follows the proof of Proposition 2.2 in Chapter 2.

We now present a sufficient statistic and a structured result for v^k after the following definition. Let g^k be the solution to the equation

$$g^k(d(t,\tau), \pi^L, \pi^F, Q)$$
$$= R^k(d(t,\tau), \pi^L, \pi^F) + \beta \sum_{\varsigma(t+1)} P(\varsigma(t+1)|d(t,\tau), \pi^L, \pi^F, Q)$$
$$\times g^k(\{\varsigma(t+1), d(t,\tau-1)\}, \pi^L, \pi^F, Q),$$

where

$$R^k(d(t,\tau), \pi^L, \pi^F) = \sum_{a(t)} r^k(s(t), a(t))P(a(t)|d(t,\tau)).$$

Proposition 3.2. *For policies π^L and π^F and observation matrix Q,*

$$v^k(\pi^L, \pi^F, Q)(d^k(t)) = v^k(\pi^L, \pi^F, Q)(d^k(t, \tau), y^k(t))$$
$$= \sum_{d^l(t,\tau)} P(d^l(t, \tau)|d^k(t))g^k(d(t, \tau), \pi^L, \pi^F, Q), l \neq k.$$

Proof follows from Lemma 1 in Ortiz, Erera and White (2013) and Proposition 2.2 in Chapter 2.

3.4.3 Partition of Observation Matrices

Let $K(\pi^L, \pi^F) = \{Q \in \mathscr{Q} : v^F(\pi^L, \pi^F, Q)(d^F(0)) \geq v^F(\pi^L, \rho^F, Q)(d^F(0)), \forall \rho^F \in \Pi^F\}$. Thus, $K(\pi^L, \pi^F)$ is the set of all matrices $Q = \{P(z^L|s^F)\}$ such that given π^L, the follower will select policy π^F (i.e., $\pi^*(\pi^L, Q) = \pi^F$). We assume that the policies in Π^F have been selected so that if $\pi^F, \rho^F \in \Pi^F$ and $\pi^F \neq \rho^F$, then $K(\pi^L, \pi^F) \cap K(\pi^L, \rho^F)$ may be non-empty but has (Lebesgue) measure zero. Thus, $\{K(\pi^L, \rho^F) : \rho^F \in \Pi^F\}$ is a finite partition of $\mathscr{Q}$ (permitting non-empty intersections when equalities occur); hence, there exists at least one element of this partition that has a non-empty interior. We assume that the follower selects π^F in order to maximize its criterion value. Thus, if Q is a member of only $K(\pi^L, \pi^F)$, then the follower will select π^F in response to the leader selecting π^L. If $Q \in K(\pi^L, \pi^F) \cap K(\pi^L, \rho^F)$, then either π^F or ρ^F may be selected. We assume that the leader has complete knowledge of $\{K(\pi^L, \rho^F) : \rho^F \in \Pi^F\}$. We remark that $K(\pi^L, \pi^F)$ for each $\pi^F \in \Pi^F$ is compact. Figure 3.1 is an illustration of a partition over $\mathscr{Q} = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$, where different colored regions correspond to different response policies ρ^F.

3.4.4 Changing Q Within A Partition Element

We now examine the impact on $v^L(\pi^L, \pi^F, Q)(d^L(t))$ of changing Q to Q' when $Q, Q' \in K(\pi^L, \pi^F)$. We begin by extending Theorem 3.1 and Corollary 3.1 to the POMG. We

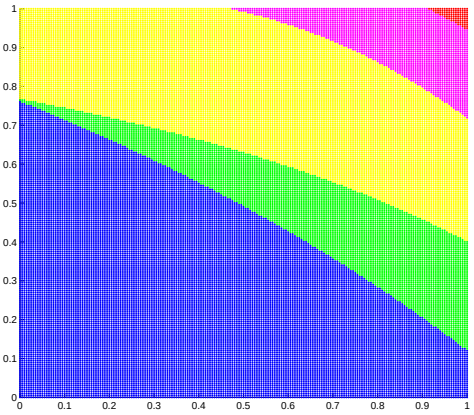


Figure 3.1: An Example of a Partition Over $\mathscr{Q} = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$.

then show that $v^L(\pi^L, \pi^F, Q)(d^L(t))$ is Lipschitz continuous on $K(\pi^L, \pi^F)$. Finally, we present conditions that guarantee two sufficiently close observation matrices are members of the same element of the partition $\{K(\pi^L, \rho^F) : \rho^F \in \Pi^F\}$.

Corollary 3.2. *Assume $Q, Q' \in K(\pi^L, \pi^F)$ and there is a $R \in \mathscr{Q}$ such that $Q'R = Q$.*

(i) If $v^L(\pi^L, \pi^F, Q)(d^L(t, \tau), y^L(t))$ is concave in $y^L(t)$ for $d^L(t, \tau)$, then

$$v^L(\pi^L, \pi^F, Q')(d^L(t, \tau), y^L(t)) \geq v^L(\pi^L, \pi^F, Q)(d^L(t, \tau), y^L(t)).$$

(ii) Let π^L be such that

$$v^L(\pi^L, \pi^F, Q)(d^L(t, \tau), y^L(t)) \geq v^L(\rho^L, \pi^*(\rho^L, Q), Q)(d^L(t, \tau), y^L(t))$$

for all $\rho^L \in \Pi^L$. Then,

$$v^L(\pi^L, \pi^F, Q')(d^L(t, \tau), y^L(t)) \geq v^L(\pi^L, \pi^F, Q)(d^L(t, \tau), y^L(t)).$$

Proof. (i) follows directly from Theorem 3.1 and Proposition 3.1. It is sufficient to show that $v^L(\pi^L, \pi^F, Q)(d^L(t, \tau), y^L(t))$ is concave in $y^L(t)$ for (ii) to hold. It follows

from Proposition 3.2 that

$$v^L(\pi^L, \pi^F, Q)(d^L(t,\tau), y^L(t))$$
$$= \max_{\rho^L} \sum_{d^F(t,\tau)} P(d^F(t,\tau)|d^L(t)) g^L(d(t,\tau), \rho^L, \pi^*(\rho^L, Q), Q),$$

which is concave in $y^L(t)$.

□

Thus, within an element of $\{K(\pi^L, \rho^F) : \rho^F \in \Pi^F\}$, Corollary 3.2 gives conditions that insure improved observation quality (based on Definition 3.1) will improve the leader's value function. We remark that Corollary 3.2 essentially extends the notion of Q-adaptivity to a pair of policies (π^L, π^F) when $Q, Q' \in K(\pi^L, \pi^F)$.

We note that Π^L and Π^F may not contain a π^L and a $\pi^*(\pi^L, Q)$ such that these equilibrium conditions hold for all initial conditions. Determining conditions that guarantee the existence of such policies is a future research topic.

We now show that for all $d^L(t)$ and for any pair of finite-memory policies (π^L, π^F), $v^L(\pi^L, \pi^F, Q)(d^L(t))$ is Lipschitz continuous in Q on $K(\pi^L, \pi^F)$.

Proposition 3.3. *For all $(d^L(t,\tau), y^L(t))$ and for any pair of finite-memory policies (π^L, π^F), $v^L(\pi^L, \pi^F, Q)(d^L(t,\tau), y^L(t))$ is Lipschitz continuous in Q on $K(\pi^L, \pi^F)$.*

Proof. For any scalar valued function v dependent on $(d(t,\tau))$, define

$$[Hv](d(t,\tau)) = R^L(d(t,\tau), \pi^L, \pi^F)$$
$$+ \beta \sum_{\varsigma(t+1)} P(\varsigma(t+1)|d(t,\tau), \pi^L, \pi^F, Q) \times v(\{\varsigma(t+1), d(t,\tau-1)\}).$$

Define H' identically to H but replace Q by Q', where we note:

$$P(\varsigma(t+1)|d(t,\tau),\pi^L,\pi^F,Q)$$
$$=\sum_{a(t)} P(a(t)|d(t,\tau))P(z^L(t+1)|s^F(t+1))P(z^F(t+1),s(t+1)|s(t),a(t)).$$

Let g and g' be the fixed points of H and H', respectively (Existence and uniqueness of these fixed points are assured by Theorem 6.2.3 in Puterman 1994). Then,

$$g(d(t,\tau))-g'(d(t,\tau))=[Hg](d(t,\tau))-[H'g'](d(t,\tau))-[Hg'](d(t,\tau))+[Hg'](d(t,\tau))$$
$$=\beta\sum_{a(t)} P(a(t)|d(t,\tau))\times X,$$

where

$$X=\sum_{z(t+1)}\sum_{s(t+1)} P(z(t+1),s(t+1)|s(t),a(t))$$
$$\times\{g(\{\varsigma(t+1),d(t,\tau-1)\})-g'(\{\varsigma(t+1),d(t,\tau-1)\})\}$$
$$+\sum_{z(t+1)}\sum_{s(t+1)}[P(z^L(t+1)|s^F(t+1))-P'(z^L(t+1)|s^F(t+1))]P(z^F(t+1),s(t+1)|s(t),a(t))$$
$$\times g'(\{\varsigma(t+1),d(t,\tau-1)\})$$

Note, $||g'||\leq\frac{M}{1-\beta}$, where $M=\max_s\max_a r^L(s,a)$. Then, it is straightforward to show that

$$||g-g'||\leq\beta||g-g'||+\beta||Q-Q'||\frac{M}{1-\beta}$$

and hence,

$$||g-g'||\leq\frac{\beta M}{(1-\beta)^2}||Q-Q'||.$$

□

Thus, as long as we remain in one of the elements of the partition $\{K(\pi^L,\rho^F):\rho^F\in\Pi^F\}$, $v^L(\pi^L,\pi^F,Q)(d^L(t,\tau),y^L(t))$ is Lipschitz continuous in Q.

We now present conditions that guarantee that two observation matrices are members of the same partition element. Consider the following definitions:

(i) for any vector g, $||g|| = \max_s |g(s)|$;

(ii) for any matrix Q, $||Q|| = \max_j \sum_{z \in Z} |q_{jz}|$.

Define $B(\rho^F) = v^F(\pi^L, \pi^F, Q')(d^F(0)) - v^F(\pi^L, \rho^F, Q')(d^F(0))$ for a given $d^F(0)$, and $b = \min\{B(\rho^F) : \rho^F \in \Pi^F, \rho^F \neq \pi^F\} \geq 0$.

Proposition 3.4. *Assume $Q' \in K(\pi^L, \pi^F)$. If Q is such that $||Q - Q'|| \leq \frac{b(1-\beta)^2}{2\beta M}$ where $M = \max_s \max_a r^L(s, a)$, then $Q \in K(\pi^L, \pi^F)$.*

Proof. If Q is such that

$$v^F(\pi^L, \pi^F, Q)(d^F(0)) - v^F(\pi^L, \rho^F, Q)(d^F(0)) \geq 0$$

for all $\rho^F \neq \pi^F, \rho^F, \pi^F \in \Pi^F$ and given $d^F(0)$, then $Q \in K(\pi^L, \pi^F)$. Note

$$\begin{aligned}
&v^F(\pi^L, \pi^F, Q)(d^F(0)) - v^F(\pi^L, \rho^F, Q)(d^F(0)) \\
&= v^F(\pi^L, \pi^F, Q)(d^F(0)) - v^F(\pi^L, \pi^F, Q')(d^F(0)) + v^F(\pi^L, \rho^F, Q')(d^F(0)) \\
&- v^F(\pi^L, \rho^F, Q)(d^F(0)) + v^F(\pi^L, \pi^F, Q')(d^F(0)) - v^F(\pi^L, \rho^F, Q')(d^F(0)) \\
&\geq -\frac{2\beta M}{(1-\beta)^2}||Q - Q'|| + b,
\end{aligned}$$

where the inequality follows from Proposition 3.3 and assumptions (ii) and (iii). The result follows from the fact that $b - \frac{2\beta M}{(1-\beta)^2}||Q - Q'|| \geq 0$ if and only if $||Q - Q'|| \leq \frac{b(1-\beta)^2}{2\beta M}$.

□

Hence, as long as observation matrices Q and Q' are close enough, they are in the same partition element which shares the same response policy. Assume current leader policy favors more accurate observation quality, Proposition 3.4 indicates how much the observation quality can improve safely without the follower changing its policy. Section 3.4.5 will show that when the observation quality is changed large enough so that the follower changes its policy, discontinuities will occur and these discontinuities can be beneficial or not beneficial to the leader.

3.4.5 Changing Q Across Partition Elements

We now examine the impact of changing Q to Q' on $v^L(\pi^L, \pi^F, Q)(d^L(t,\tau), y^L(t))$ when $Q \in K(\pi^L, \pi^F)$, $Q' \in K(\pi^L, \rho^F)$, and $\pi^F \neq \rho^F$. Without loss of generality, assume $Q^* \in K(\pi^L, \pi^F) \cap K(\pi^L, \rho^F)$, $\{Q_n\}$ is a sequence in $K(\pi^L, \pi^F)$ that converges to Q^*, and $\{Q'_n\}$ is a sequence in $K(\pi^L, \rho^F)$ that converges to Q^*. Then, from the Proposition 3.3 and the compactness of $K(\pi^L, \pi^F)$ and $K(\pi^L, \rho^F)$,

$$\lim_{n\to\infty} v^L(\pi^L, \pi^F, Q_n)(d^L(0)) = v^L(\pi^L, \pi^F, Q^*)(d^L(0))$$

$$\lim_{n\to\infty} v^L(\pi^L, \rho^F, Q'_n)(d^L(0)) = v^L(\pi^L, \rho^F, Q^*)(d^L(0)).$$

However, there is no guarantee that $v^L(\pi^L, \pi^F, Q^*)(d^L(0)) = v^L(\pi^L, \rho^F, Q^*)(d^L(0))$, and hence, at a boundary there can be discontinuities. Figure 3.2 presents a 3-dimensional view of possible discontinuities at the boundaries of the partition elements over $\mathscr{Q} = \{(x,y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$, where different colored regions correspond to different response policies ρ^F.

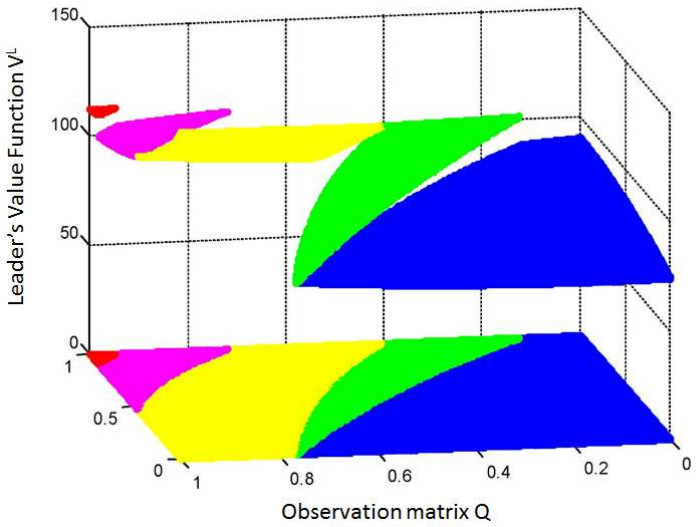


Figure 3.2: An Example of Discontinuities at the Boundaries of the Partition Elements Over $\mathscr{Q} = \{(x,y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$

We now show by example a variety of ways that v^L can depend on Q as Q moves across

boundaries in the partition $\{K(\pi^L, \pi^F), \pi^F \in \Pi^F\}$. In order to reduce computational complexity, we assume that $a^L(t)$ depends only on $\{s^L(t), z^L(t)\}$ and that the follower can completely observe the leader's information $\{s^L(t), z^L(t)\}$. Parameter values for all examples are presented in Appendix.

Example 3.1. *Let* $|S^L| = |A^L| = |Z^L| = |S^F| = |A^F| = 2$. *Let* $Q' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ *and* $R(\epsilon) = \begin{bmatrix} 1-\epsilon & \epsilon \\ \epsilon & 1-\epsilon \end{bmatrix}$ *where* $\epsilon > 0$ *and is given. Define* $Q(x) = Q'R^x$, *then* $Q(x)$ *is a stochastic matrix for all* $x \geq 0$ *and* $Q(x_1)$ *is at least as informative as* $Q(x_2)$ *if* $0 \leq x_1 \leq x_2$.

Note, the Markov chain constructed by R *is aperiodic and irreducible, hence there exists a unique* Q^* *such that* $Q^* = \lim_{x\to\infty} Q'R^x$ *and* $Q^* = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$, *corresponding to the completely unobserved case. In addition, the second largest eigenvalue value of* R *is* $1 - 2\epsilon$; *hence* $Q(x)$ *converges to* Q^* *faster as* ϵ *increases.*

Figure 3.3 shows the changes in the leader's value function for a given leader's policy as observation quality degrades for six examples. In these examples, only $r^L(s(t), a(t))$ *was varied;* $P(z(t+1), s(t+1)|s(t), a(t))$ *and* $r^F(s(t), a(t))$ *remained unchanged. A discontinuity can occur when the change in* $Q(x)$ *is great enough to cause a change in the follower's policy. Note that the leader completely observes the follower when* $x = 0$, *whereas when* x *is large (*$x \geq 20$ *in these examples), the leader receives no information about the follower from observations. Regarding the examples in Figure 3.3, all six have three discontinuities. As* x *increases, we note discontinuities that produce abrupt increases or decreases in the leader's value function, and between the discontinuities we note monotone increasing or decreasing value function performance.*

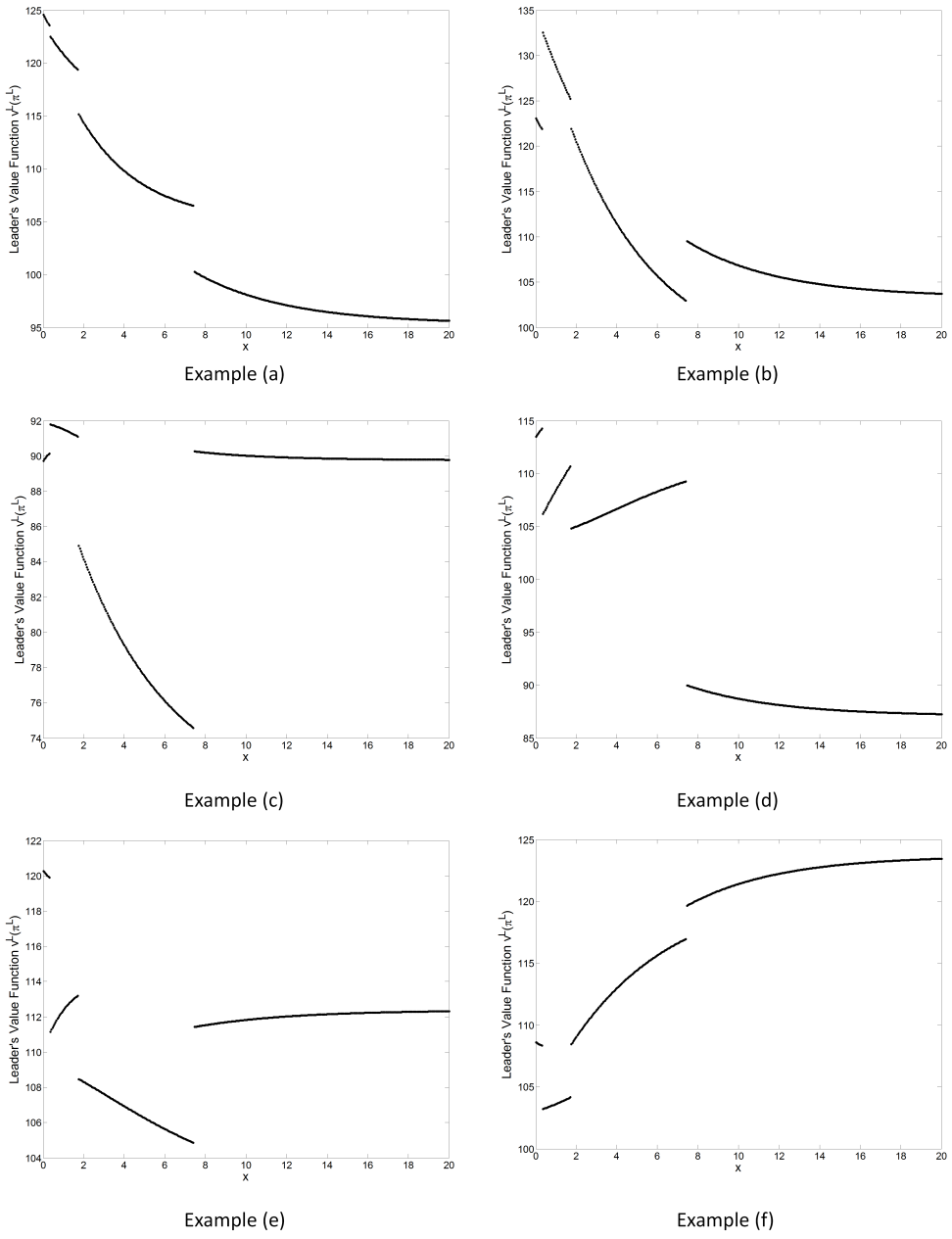


Figure 3.3: A Variety of Changes to the Leader's Value Function as Observation Quality Degrades

The examples have shown that if the leader's observation matrix changes from $Q \in K(\pi^L, \pi^F)$ to $Q' \in K(\pi^L, \rho^F), \pi^F \neq \rho^F$, then the leader's value function may experience an abrupt change of value due to discontinuities that can occur at partition boundaries and that these changes can be favorable or unfavorable. We now present

sufficient conditions that guarantee a favorable change.

Let $N = (|Z^L||Z^F||S^L||S^F||A^L||A^F|)^\tau$, and assume μ is a one-to-one, onto mapping from $\{d(t,\tau)\}$, the set of all $d(t,\tau)$, to $\{1,2,...,N\}$. Thus, μ totally orders $\{d(t,\tau)\}$.

A function $f : \{d(t,\tau)\} \to R$ is said to be isotone (with respect to μ) if and only if $\mu(d(t,\tau)) \leq \mu(d'(t,\tau))$ implies $f(d(t,\tau)) \leq f(d'(t,\tau))$.

For any $\pi' \in \Pi^F$ and $Q \in \mathscr{Q}$, let

$$q(k|d(t,\tau),Q,\pi^L,\pi') = \sum P(\varsigma(t+1)|d(t,\tau),Q,\pi^L,\pi'),$$

where the sum is over all $\varsigma(t+1)$ such that $\mu(\{\varsigma(t+1),d(t,\tau-1)\}) \geq k$.

Lemma 3.1. *Assume:*

(i) $R^L(d(t,\tau),\pi^L,\pi^F)$ *is isotone in* $d(t,\tau)$,

(ii) $q(k|d(t,\tau),Q,\pi^L,\pi^F)$ *is isotone in* $d(t,\tau)$ *for all* k,

Then, $g^L(d(t,\tau),Q,\pi^L,\pi^F)$ *is isotone in* $d(t,\tau)$.

The proof follows from Proposition 4.7.3 (Puterman, p. 106) and a standard limit procedure.

Proposition 3.5. *Assume:*

(i) $Q^* \in K(\pi^L,\pi^F) \cap K(\pi^L,\rho^F), \pi^F \neq \rho^F$,

(ii) $R^L(d(t,\tau),\pi^L,\pi')$ *is isotone in* $d(t,\tau)$ *for* $\pi' \in \{\pi^F,\rho^F\}$,

(iii) $q(k|d(t,\tau),Q^*,\pi^L,\pi')$ *is isotone in* $d(t,\tau)$ *for all* k *for* $\pi' \in \{\pi^F,\rho^F\}$,

(iv) $R^L(d(t,\tau),\pi^L,\rho^F) \geq R^L(d(t,\tau),\pi^L,\pi^F)$ *for all* $d(t,\tau)$,

(v) $q(k|d(t,\tau),Q^*,\pi^L,\rho^F) \geq q(k|d(t,\tau),Q^*,\pi^L,\pi^F)$ *for all* k *and all* $d(t,\tau)$.

Then, $g^L(d(t,\tau),Q^*,\pi^L,\rho^F) \geq g^L(d(t,\tau),Q^*,\pi^L,\pi^F)$ *for all* $d(t,\tau)$*, and hence* $v^L(\pi^L,\rho^F,Q^*)(d^L(0)) \geq v^L(\pi^L,\pi^F,Q^*)(d^L(0))$.

Proof. Lemma 3.1 guarantees that $g^L(d^L(t,\tau),Q^*,\pi^L,\pi')$ for $\pi' \in \{\pi^F,\rho^F\}$ is isotone in $d(t,\tau)$. It follows from Lemma 4.7.2 (Puterman, p. 106) that

$$\begin{aligned}&\sum_{\varsigma(t+1)} P(\varsigma(t+1)|d(t,\tau),Q^*,\pi^L,\rho^F) \times g^L(d(t,\tau)|Q^*,\pi^L,\pi^F)\\ &\geq \sum_{\varsigma(t+1)} P(\varsigma(t+1)|d(t,\tau),Q^*,\pi^L,\pi^F) \times g^L(d(t,\tau)|Q^*,\pi^L,\pi^F).\end{aligned}$$

Thus,,

$$\begin{aligned}&g^L(d(t,\tau),Q^*,\pi^L,\pi^F)\\ &\leq R^L(d(t,\tau),\pi^L,\rho^F) + \beta \sum_{\varsigma(t+1)} P(\varsigma(t+1)|Q^*,\pi^L,\rho^F) \times g^L(d(t,\tau),Q^*,\pi^L,\pi^F).\end{aligned} \tag{3.5}$$

Let

$$\begin{aligned}&[Hv](d(t,\tau))\\ &= R^L(d(t,\tau),\pi^L,\rho^F) + \beta \sum_{\varsigma(t+1)} P(\varsigma(t+1)|d(t,\tau),Q^*,\pi^L,\rho^F)\\ &\times v(\{\varsigma(t+1),d(t,\tau-1)\}).\end{aligned}$$

Define the sequence $\{v^n\}$ as $v^{n+1} = Hv^n$, where $v^0(d(t,\tau)) = g^L(d(t,\tau),Q^*,\pi^L,\pi^F)$. We remark that $\lim_{n\to\infty} ||v^n - v^*|| = 0$, where $v^*(d(t,\tau)) = g^L(d(t,\tau),Q^*,\pi^L,\rho^F)$. It is straightforward to show that $v \leq v'$ implies $Hv \leq Hv'$. Equation (3.5) has shown $v^0 \leq v^1$. Lemma 3.1 guarantees that v^n is isotone in $d(t,\tau)$ for $n \geq 1$. Hence, by induction, $v^n \leq v^{n+1}$ and therefore $v^n \leq v^*$ for all n. Thus, $g^L(d(t,\tau),Q^*,\pi^L,\pi^F) \leq g^L(d(t,\tau),Q^*,\pi^L,\rho^F)$ for all $d(t,\tau)$ and hence $v^L(\pi^L,\pi^F,Q^*)(d^L(0)) \leq v^L(\pi^L,\rho^F,Q^*)(d^L(0))$.

□

Proposition 3.5 presents conditions involving both $R^L(d(t,\tau),\pi^L,\pi')$ and $P(\varsigma(t+1)|d(t,\tau),Q^*,\pi^L,\pi')$ that suggest a change in observation quality from $Q \in K(\pi^L,\pi^F)$ to $Q' \in K(\pi^L,\rho^F)$, where Q and Q' are both close to $Q^* \in K(\pi^L,\pi^F) \cap K(\pi^L,\rho^F)$, will improve the leader's performance.

It is easily shown that the result in Proposition 3.5 holds if

$$\max_{d(t,\tau)} R^L(d(t,\tau),\pi^L,\pi^F) \leq \min_{d(t,\tau)} R^L(d(t,\tau),\pi^L,\rho^F), \tag{3.6}$$

which does not require assumptions on $P(\varsigma(t+1)|d(t,\tau),Q^*,\pi^L,\pi')$. We remark, however, that (3.6) is considerably stronger than Assumption (iv) in Proposition 3.5 and hence may be more difficult to be satisfied than the assumptions in Proposition 3.5.

Example 3.2. *Consider the problem in Example 3.1, assuming the leader's information can be only partially observed by the follower and* $P(d^L(0)|d^F(0)) = 1$. *Figure 3.4 shows that a favorable change in leader's value function can occur at partition boundaries when the assumptions in Proposition 3.5 are satisfied.*

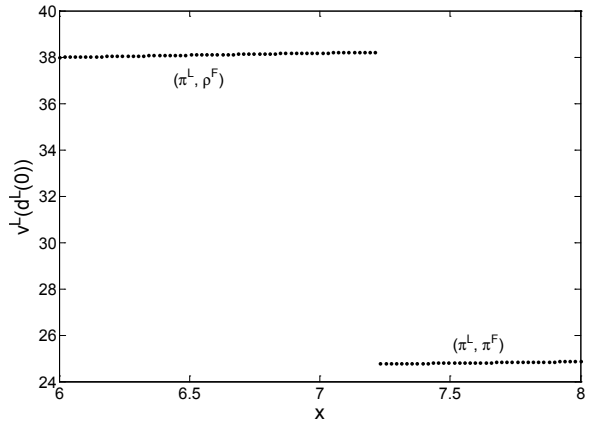


Figure 3.4: Favorable Change in Leader's Value Function Across the Boundary

We now examine the situation where the two agents are collaborative (share the same objective; i.e. $r^L = r^F$) and at least the follower has complete knowledge of the leader's initial finite-memory state of knowledge ($d^L(0,\tau)$). Under these conditions there will not be a discontinuity in crossing a boundary of the partition $\{K(\pi^L, \pi^F) : \pi^F \in \Pi^F\}$.

Proposition 3.6. *For* $d^L(0) = \{d^L(0,\tau), y^L(0)\}$ *and* $d^F(0) = \{d^F(0,\tau), y^F(0)\}$, *assume:*

(i) $P(d^L(0,\tau)|d^F(0)) = 1$,

(ii) $Q^* \in K(\pi^L, \pi^F) \cap K(\pi^L, \rho^F), \pi^F \neq \rho^F$,

(iii) $r^L = r^F$,

Then, $v^L(\pi^L, \pi^F, Q^*)(d^L(0)) = v^L(\pi^L, \rho^F, Q^*)(d^L(0))$.

Proof. Assumption (i) implies that $v^F(\pi^L, \pi', Q^*)(d^F(0)) = g^F(d(0,\tau), Q^*, \pi^L, \pi')$ for $\pi' \in \{\pi^F, \rho^F\}$. Assumption (ii) implies $v^F(\pi^L, \pi^F, Q^*)(d^F(0)) = v^F(\pi^L, \rho^F, Q^*)(d^F(0))$. Hence, $g^F(d(0,\tau), \pi^L, \pi^F) = g^F(d(0,\tau), \pi^L, \rho^F)$. Assumption (iii) implies $g^L = g^F$. □

We now provide several illustrative examples under the conditions assumed in Proposition 3.6.

Example 3.3. *Assume the problem in Example 3.1 under the assumptions of Proposition 3.6. Figure 3.5 shows the changes to the leader's value function as observation quality degrades for four randomly generated examples. All discontinuity points disappear, and the value function of the leader* v^L *is continuous with respect to observation matrix* Q. *However, the slope of the value function* v^L *and its sign can still be negative or positive due to the changes of the best response policy* ρ^F.

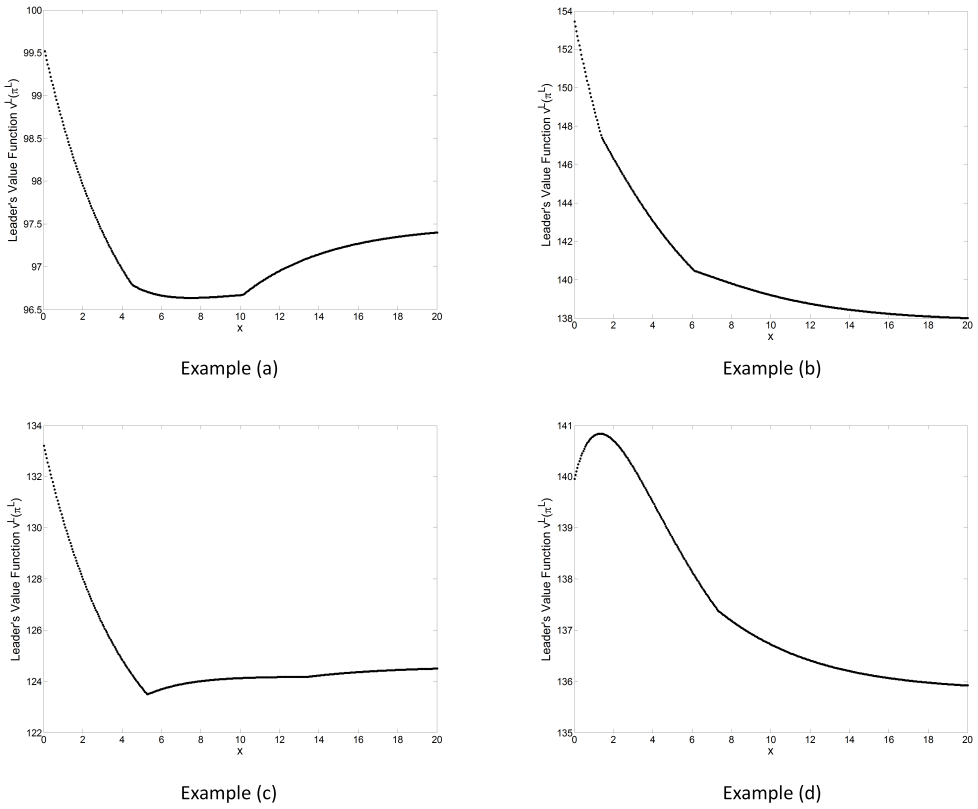


Figure 3.5: Changes to the leader's value function as observation quality degrades under conditions of Proposition 3.6

The implications of these illustrative examples are that even if the two agents are totally collaborative (i.e., $r^L = r^F$) and initially at least the follower has complete visibility of the leader's state (i.e., $y^F(0)$ identifies $d^L(0, \tau)$ with probability 1), improved observation quality for the leader may or may not improve system performance, before or after a boundary is crossed. These results indicate that greater visibility between even collaborative agents may not result in improved system performance. Determining conditions under which greater visibility between collaborative agents will result in improved system performance is a topic for future research.

3.5 Conclusions

We have examined how changes in the accuracy of the leader's observation of the follower can affect the leader's value function. We have given conditions that insure improved observation quality improve the leader's value function, assuming the changes in observation quality do not cause the follower to change its policy. We demonstrated that when changes in observation quality cause the follower to change its policy, discontinuities in the leader's value function can result, as a function of observation quality, and that these discontinuities can be beneficial or not beneficial to the leader. We showed that when the two agents are collaborative, i.e., share the same reward structure, and the follower has complete visibility of the leader's initial conditions, discontinuities in the leader's value function do not occur. However, whether or not the agents in the POMG are collaborative, improved quality of the leader's observations of the follower do not necessarily lead to improved leader performance.

This research represents an initial investigation into the impact of observation quality on performance for the POMG under very specific assumptions (there are two agents, a leader and a follower, and the follower selects its policy with complete knowledge of the leader's policy selection) and with focus on how the leader's observation quality of the follower impacts the leader's value function. Future directions for research on the interplay between observation quality and control in the context of the POMG appear numerous.

3.6 References

[1] Bassan, B., Gossner, O., Scarsini, M., and Zamir, S. (2003) Positive value of information in games, *Internal Journal of Game Theory*, 32: 17 - 31.

[2] Bertsekas, D. P., (1976) Dynamic programming and stochastic control, Academic Press, New York.

[3] Bernstein, D. S., Givan, R., Immerman, N., and Zilberstein, S. (2002) The complexity of decentralized control of Markov decision processes, *Mathematics of Operations Research*, 27(4): 819 - 840.

[4] Bier V. M., Oliveros, S., and Samuelson, L. (2007) Choosing what to protect: Strategic defensive allocation against an unknown attacker, *Journal of Public Economic Theory*, 9(4): 563 - 587.

[5] Cassandra, A. R., Kaelbling, L. P., and Littman, M. L. (1994) Acting optimally in partially observable stochastic domains , in*Proceedings Twelfth National Conference on Artificial Intelligence (AAAI-94)*, Seattle, WA, 1023 - 1028.

[6] Cassandra, A. R., Littman, M. L. and Zhang, N. L. (1997) Incremental pruning: a simple, fast, exact method for partially observable Markov decision processes, in *Proceedings Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-97)*, Morgan Kaufmann, San Francisco, CA, 54 - 61.

[7] Chu, W. H. J, and Lee, C. C. (2006) Strategic information sharing in a supply chain, *European Journal of Operational Reserch*, 174, 1567 - 1579.

[8] Ezell, B. C., Bennett, S. P., von Winterfeldt D., Sokolowski, J. and Collins, A. J. (2010) Probabilistic risk analysis and terrorism risk, *Risk Analysis*, 30(4): 575-589.

[9] Hansen, E. A., Bernstein, D. S., and Zilberstein, S. (2004) Dynamic programming for partially observable stochastic games, in *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, 709 - 715, San Jose, California.

[10] Kamien, M. I., Tauman, Y., and Zamir, S. (1990) On the value of information in a strategic conflict, *Games and Economic Behavior* , 2: 129 - 153.

[11] Kumar, A., and Zilberstein, S. (2009) Dynamic programming approximations for partially observable stochastic games, In *Proceedings of the Twenty-second International FLAIRS Conference*, 547 - 552, Sanibel Island, Florida.

[12] Lehrer, E., and Rosenberg, D. (2006) What restrictions do Bayesian games impose on the value of information?, *Journal of Mathematical Economics*, 42: 343 - 357.

[13] Lehrer, E., and Rosenberg, D. (2010) A note on the evaluation of information in zero-sum repeated games, *Journal of Mathematical Economics*, 46: 393 - 399.

[14] Leng, M. M., Parlar, M. (2009) Allocation of cost savings in a three-level supply chain with demand information sharing: A cooperate-game approach, *Operations Research*, 57(1): 200 - 213.

[15] Li, L. (2002) Information sharing in a supply chain with horizontal competition, *Management Science*, 48(9): 1196 - 1212.

[16] Lin, A. Z.-Z.,Bean, J., and White, C. C. (1998) Genetic algorithm heuristics for finite horizon partially observed Markov decision problems, *Technical Report*, University of Michigan, Ann Arbor.

[17] Lin, A. Z.-Z.,Bean, J., and White, C. C. (2004) A hybrid genetic/optimization algorithm for finite horizon partially observed Markov decision processes, *Journal on Computing*, 16(1): 27 - 38.

[18] Littman, M. L. (1994) The Witness algorithm: solving partially observable Markov decision processes, Brown University, Department of Computer Science, *Technical Report*, CS-94-40.

[19] Lovejoy, W. S. (1991) A survey of algorithmic methods for partially observed Markov decision process, *Annals of Operations Research*, 28(1): 47 - 65.

[20] Meuleau, N., Peshkin, L., Kim, K., and Kaelbling, L. P. (1999) Learning finite-state controllers for partially observable environments, in *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, 427 - 436.

[21] Meyer, B. D., Lehrer, E., and Rosenberg, D. (2010) Evaluating information in zero-sum games with incomplete information on both sides, *Mathematics of Operations Research*, 35(4), 851 - 863.

[22] Monahan, G. E. (1982) A survey of partially observable Markov decision processes: Theory, models, and algorithms, *Management Science*, 28: 1 - 16.

[23] Ortiz, O. L., Erera, A. L., White, C. C. (2013) State observation accuracy and finite-memory policy performance, *Operations Research Letters*, 41: 477 - 481.

[24] Platzman, L. K. (1977) Finite memory estimation and control of finite probabilistic systems, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

[25] Platzman, L. K. (1980) Optimal infinite-horizon undiscounted control of finite probabilistic systems, *SIAM Journal on Control and Optimization*, 18: 362 - 380.

[26] Poupart, P. and Boutilier, C. (2004) Bounded finite state controllers, *Advances in Neural Information Processing Systems*, 16, MIT Press, Cambridge, MA.

[27] Puterman, M. L. (1994) Markov decision processes: discrete dynamic programming, New York: J Wiley & Sons.

[28] Rabinovich, Z., Goldman, C. V., and Rosenschein, J. S. (2003) The complexity of multiagent systems: the price of silence, In *Proceedings of the Second International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 1102 - 1103, Melbourne, Australia.

[29] Shapley, L. S. (1953) Stochastic games, *Proceedings of the National Academy of Sciences of the U. S. A.*, 39: 1095 - 1100.

[30] Smallwood, R. D., and Sondik, E. J. (1973) The optimal control of partially observable Markov decision processes over a finite horizon, *Operations Research*, 21: 1071 - 1088.

[31] Sondik, E. J. (1978) The optimal control of partially observable Markov processes over the infinite horizon: discounted costs, *Operations Research*, 26: 282 - 304.

[32] Wakker, P., (1988) Nonexpected utility as aversion of information, *Journal of Behavioral Decision Making*, 1: 169 - 175.

[33] White, C. C., and Harrington, D. P. (1980) Application of Jensen's inequality to adaptive suboptimal design, *Journal of Optimization Theory and Application*, 32: 89 - 99.

[34] White, C. C., and Scherer, W. T. (1989) Solution procedures for partially observed Markov decision processes, *Operations Research*, 37: 791 - 797, 1989.

[35] White, C. C. (1991) A survey of solution techniques for the partially observed Markov decision process, *Annals of Operations Research*, 32: 215 - 230.

[36] White, C. C., and Scherer, W. T. (1994) Finite-memory suboptimal design for partially observed Markov decision processes, *Operations Research*, 42: 439 - 455.

[37] Zhang, H. (2010) Partially observable Markov decision processes: a geometric technique and analysis, *Operations Research*, 58: 214 - 228.

[38] Zhuang, J., Bier, V. M. and Alagoz, O. (2010) Modeling secrecy and deception in a multiple-period attacker-defender signaling game, *European Journal of Operational Research*, 203: 409 - 418.

[39] Zhuang, J., and Bier, V. M. (2010) Reasons for secrecy and deception in homeland-security resource allocation, *Risk Analysis*, 30(12): 1737 - 1743.

CHAPTER IV

RISK ASSESSMENT OF DELIBERATE CONTAMINATION OF FOOD PRODUCTION FACILITIES

4.1 Introduction

The deliberate contamination of food is recognized as a major global public health threat, and food defense is one of the 17 national critical sectors (DHS, 2007) in the U.S.. The U. S. Food and Drug Administration (FDA) budget requests for fiscal year 2014 have reached $4.7 billion to build a strong and reliable food system (FDA, 2013). Deliberate food contamination can have disastrous impact from both social and economic perspectives. One of the most well-known examples of deliberate food contamination is the outbreak of Salmonella Typhimurium that occured in Dalles, Oregon, in September and October 1984 and sickened 751 people, 45 of whom were hospitalized (USCB, 2011).

FDA (USDHHS, 2007) has recommended that food industry operators review their current procedures and controls, assess potential threats and the effectiveness of preventive measures, and implement enhanced preventive measures. Existing work has been focused on estimating the probability of an attack at a given target and the consequence of an attack based on statistical data (McGill, Ayyub and Kaminskiy, 2007). Such analysis does not account the fact that both the attacker and the defender are intelligent and adaptive and can take action based on possibly inaccurate information about the other agent. Hence, these probabilities can change as the current situation changes and may be affected by the accuracy of the observations. To overcome these

difficulties, we use a game with multiple decision epochs to describe the sequential interaction between the defender and the attacker, extending the current literature (Bier, Oliveros and Samuelson, 2007; Bier et al., 2008; Hao, Jin and Zhuang, 2009; Levitin and Hausken, 2010), which has focused on the allocation of resources to reduce risk based on a single decision epoch decision making assumption. We remark that using models of sequential decision making for multiple agents with inaccurate information is a relatively unexamined area in risk analysis.

We use a leader-follower two-agent, partially observed Markov game (POMG) in Chapter 2 as our model. The agents interact as follows:

- Before the game begins, the follower (the attacker) chooses its best response policy with complete knowledge of what policy the leader (the defender) has chosen.
- Once the game begins, the policies chosen by the leader and the follower determine what action to select at each decision epoch of an at most countable number of decision epochs.
- The dynamics of the system is controlled by the actions of both agents.
- Each agent's policy determines the action to take at each decision epoch, based on data that include possibly inaccurate, incomplete, and/or costly observations of the other agent.

Compared to the existing risk analysis methods, our approach has the following advantages:

- It explicitly considers the interaction between the defender and the attacker over time.

- Both the defender's and the attacker's actions are selected on the basis of data collected at current and past decision epochs.
- The data collected at each decision epoch by each agent may include possibly inaccurate, incomplete, and/or costly observations of the other agent. Hence, our risk analysis model can be used to analyze the impact of information accuracy on the risk assessment and mitigation.
- It can consider multiple objectives for the defender.

We remark that although our risk analysis approach was developed to model liquid eggs production, our approach can be applied to a much broader class of problems than those involving food supply chains.

The remainder of this chapter is organized as follows. We review the existing risk assessment literature in Section 4.2. We describe our risk analysis tool in Section 4.3, which is comprised of two components: a consequence assessment tool in Section 4.3.1 and a game theoretic optimization model in Section 4.3.2. Section 4.3.3 outlines the solution procedure. Section 4.3.4 considers four cases in terms of information asymmetry. Numerical results for a simple model of a liquid egg production facility are presented in Section 4.4. Section 4.4.1 summarizes the runtime results of our approach. Section 4.4.2 presents results for the base model. Section 4.4.3 presents a comparison of the defender's performance for four types of information asymmetry and how the defender's performance improves as data accuracy improves, representing an analysis of value of information. We also analyze how the characteristics of the defender's policies differ for each type of data accuracy asymmetry. Section 4.4.4 illustrates how system risk can be dynamic as a result of the strategic interaction between the defender and the attacker. We show how a defender's policy can

redirect the attacker's interests to less vulnerable targets and lengthen the expected time till an attack occurs. A sensitivity analysis is performed in Section 4.4.5 in order to identify particularly sensitive parameters. Conclusions are presented in Section 4.5.

4.2 Related Literature

Existing risk analysis tools for food industry include CARVER+Shock (Clark and Philpott, 2011), Food & Agriculture Sector-Criticality Assessment (FASCAT), bioterrorism terrorism risk assessment (BTRA) (NCFPD report 2007 - 2011), Operational Risk Management (ORM) (Headquarters Marine Corps, 2002), and FDA iRisk (FDACFS, 2012). These tools assume that risk is static in time. In the broader risk analysis literature, risk assessment may include uncertainties associated with risk scenarios, consequences, responses of agents, and the dynamics of the system. Probabilistic risk assessment (PRA) has been a major tool for characterizing these uncertainties and estimating structural reliability and risk (Ezell et al., 2010). References of McGill, Ayyub and Kaminskiy (2007) and Rosoff and von Winterfeldt (2010) define adversarial risk as $\sum_A P(\text{Success}|A)C(A)P(A)$, where $P(A)$ is the probability that attack A will occur; $P(\text{Success}|A)$ is the probability that attack A will be successful, given that attack A has occurred; and $C(A)$ is the consequence that results (such as measures of morbidity, mortality, economic or societal impact) if attack A is successful. Quantifying $P(A)$ is particularly challenging, and approaches for eliciting such probabilities from experts are summarized in Bedford and Cooke (2001) and Hora (2007). References of Cox (2008, 2009) and Brown and Cox (2011) have pointed out the limitations of PRA and have indicated that these limitations can produce decisions that result in an increased risk of attack. A major challenge is how to assess probabilities that are in reality not static, since an adversary may change tactics when defensive resource allocations change. Reference of Ezell et al. (2010)

has stated that the probabilities (and corresponding risk) estimated based on the current state of information can only serve as a baseline. Reference of Bompard et al. (2009) also has pointed out the importance of modeling the strategic interaction between adversaries in security analysis.

Game theory is a natural basis for modeling interacting decision makers with different information structures and/or different objectives. References of Bier, Oliveros and Samuelson (2007), Bier et al. (2008), Hao, Jin and Zhuang (2009), Levitin and Hausken (2010), and Shan and Zhuang (2013) have examined how to allocate defensive resources in order to reduce expected risk for a single decision epoch game. The resource allocation problem that considers terrorism and natural disasters simultaneously is studied in Powell (2007a, 2007b), Golany et al. (2009), and Levitin and Hausken (2009). Bi- and tri-level optimization extensive game models are used in Brown et al. (2006) to study critical infrastructure protection issues. Reference of Sandler and Siqueira (2009) showed how uncovering information about terrorists' targeting preferences could (counterintuitively) increase a government's vulnerability to a terrorist attack. In reality, it is very likely that neither the defender nor the attacker knows the rewards, objectives, actions, and beliefs of the other agent. Hence, more recent studies have developed defender-attacker models using incomplete information. For example, Bier et al. (2008) studied how to allocate a defensive budget when the defender's uncertainty about the attacker's target valuations follows a two-parameter Rayleigh distribution. The value of the defender disclosing its information to the public has also been investigated in Zhuang and Bier (2010, 2011).

The majority of the above research was based on single decision epoch games. However, such games cannot fully describe a dynamic decision-making environment and

the adaptive nature of terrorist (and defender) behavior over time. To this end, differential games have been used in Feichtinger and Novak (2008) to study the strategic interactions of Western governments and terrorist organizations over time. A two-stage repeated game in which the agents are myoptic in each period was studied in Hausken and Zhuang (2011, 2012) in order to study the timing and deterrence of terrorist attacks and defensive resource allocations. A repeated signaling game with incomplete information was used in Zhuang, Bier and Alagoz (2010), and the authors showed that the defender's objectives can be improved by secrecy and deception. The stochastic game is an ideal model for the situation where the attacker has to transition through successive states to probe a system before an attack. Reference of Bakir and Kardes (2011) studied container security using a completely observed stochastic game that can consider the uncertainties associated with the agent's states and response actions over time. However, it is very unusual in reality that both the defender and the attacker have accurate observation about the other agent's state. By incorporating the POMG, this chapter contributes to the risk analysis literature by further investigating how to assess and mitigate adversarial risk when both agents are given inaccurate observations of the other agent's state at each decision epoch.

In addition, there has been a growing literature that applies multi-objective optimization models to homeland security problems, where the defender has to consider several competing objectives simultaneously. Reference of Brown et al. (2012) has developed a multi-objective single-stage Stackelberg game played between the defender and N different types of attackers. This multi-objective security game can generate a Pareto frontier for the defender facing various types of adversarial risks. The risk analysis tool we have developed and use here considers multi-objectives for the defender in a multiple period decision making environment.

4.3 Risk Analysis Model

Our risk analysis model has two components: a game theoretic optimization model and a consequence assessment model. The game theoretic optimization model explicitly describes the strategic interaction between the defender and the attacker over time *before* an attack occurs. Once an attack occurs, the game stops and the consequence assessment model determines the expected consequence of the attack. Hence, the consequence assessment model serves as an input to the terminal reward of the game model, and the generated consequence depends on the states of the defender and the attacker at the time of the attack.

4.3.1 Consequence Assessment Model

The consequence assessment model calculates the expected impact of an attack on the system, and this expected impact depends on the structure of the system and the strategies and states of both the defender and attacker. For example, if a link or a node of a supply chain network is disrupted, the consequence assessment model should simulate the expected impact of this disruption on the entire system. We illustrate this concept using the liquid eggs production facility example we now introduce.

Figure 4.1 shows the liquid eggs production process. The shell eggs are transported to the sanitizing and grading facility. After grading, well-graded eggs (grade AA, A or B) are packed and transported to the markets, whereas off-graded eggs are transported to the breaking system. The resulting liquid eggs are collected by Collecting Vats and stored in Raw Production Tanks. The Pasteurizers reduce the amount of some toxins (e.g. population of micro-organisms) in the liquid eggs. After pasteurization, the liquid eggs flow into the Finished Product Tank and are then packaged for distribution. Figure 4.2 presents the simplified production process model found

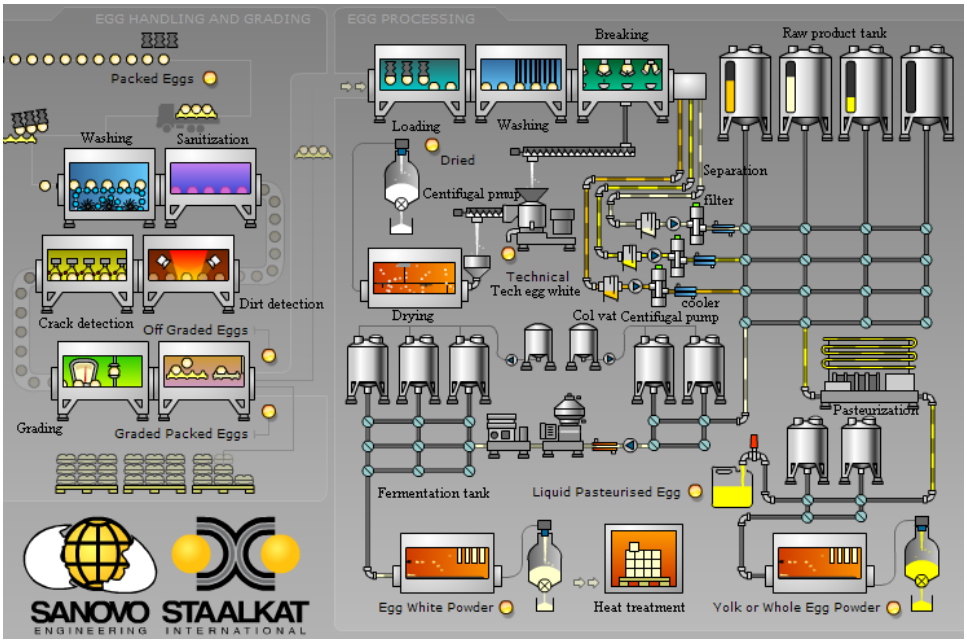


Figure 4.1: The liquid eggs processing system (Zhang, 2013)

Figure 4.2: Simplified liquid eggs processing system (Zhang, 2013)

in (Zhang, 2013).

The attacker may insert a toxin (e.g. botulinum) at various places ("targets") in this system, and the number of product units that contain a lethal amount of contaminant that leave the facility is the measure of consequence. We now define an attack scenario.

Definition 4.1. *An attack scenario is a 5-tuple* (a, m, i, t, n)*, where* $a =$ *toxin,* $m =$ *mass of the toxin inserted into the facility,* $i =$ *the target where the toxin is inserted,* $t =$ *the time when the toxin is inserted, and* $n =$ *whether the attacker can escape (0) or not (1).*

For example, the attack scenario $(a_0, m_0, i_0, t_0, 0)$ is: deliver mass $m_0 \geq 0$ of toxin a_0

to target i_0 at time t_0 and then escape successfully.

The consequence of an attack depends on the mass of toxin inserted into the facility (m), the target where the toxin is inserted (i) and the time of insertion (t). The consequence is calculated by simulating the toxin contamination process in the facility using the contamination model in Zhang (2013). In this example, we consider the worst-time scenario using the tool developed in Zhang (2013), where we calculate the worst possible consequence by determining the worst time the toxin can be inserted into the production process. Let $cons(m, i)$ be the consequence generated by inserting m amount of toxin at target i at the worst possible time during the production process.

Figure 4.3 shows that the consequence of an attack at each target is non-decreasing with increased mass of toxin. When the toxin mass is relatively high, the consequence of an attack at Collecting Vat is higher than the consequence of an attack at the Raw Product Tank, and the consequence of an attack at the Raw Product Tank is higher than the consequence of an attack at the Finished Product Tank. These statements are based on the fact that an attack upstream will affect downstream production. However, when the mass of toxin is low, the pasteurizer can reduce the effectiveness of the most of the toxin. Hence, the consequence of an attack at the Collecting Vat or Raw Product Tank will be lower than at the Finished Product Tank. When the toxin mass is sufficiently large, the entire process is contaminated; hence, the number of contaminated packages approaches an upper bound. Among the four components of the facility, the pasteurization process is the most effective in reducing or eliminating toxin effectiveness and the component most difficult to attack. Thus, we eliminate the pasteurization process as a possible target and only consider the Collecting Vat (Target 1), Raw Product Tank (Target 2) and the Finished Product Tank (Target 3) as possible targets.

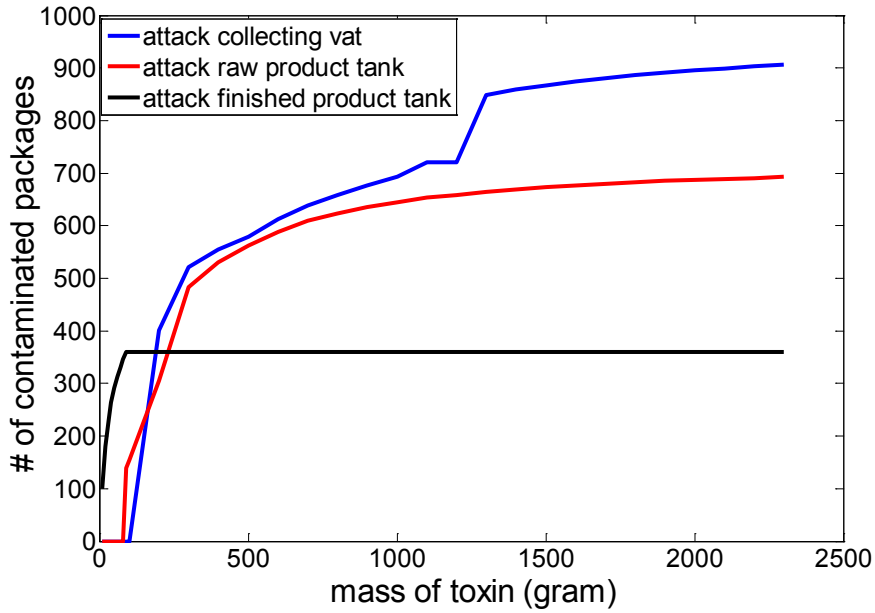


Figure 4.3: Consequence of an attack at different targets

4.3.2 Game Theoretic Optimization Model

The game theoretic optimization model is based on a partially observed Markov game, which is comprised of:

4.3.2.1 State Spaces

The defender's state space S^D The defender's state s^D represents the operation status of the system. Each target i of the system can be in one of three states $\{F, M, L\}$. State F is the high productivity, low-risk mitigation state for target i; M is the medium productivity, medium-risk mitigation state for target i; L is the low productivity, high-risk mitigation state for target i. Since there are three possible targets, the defender's state space is $S^D = \{(s_1^D, s_2^D, s_3^D), s_i^D \in \{F, M, L\}, i = 1, 2, 3\} \cup \{Att\}$, where s_i^D indicates that target i is in state s_i^D and Att indicates that an attack has occurred. For example, (M, F, L) means the Collecting Vat is in the {medium

productivity, medium-risk mitigation} state, the raw product rank is in the {high productivity, low-risk mitigation} state, and the finished production tank is in the {low productivity, high-risk mitigation} state. The size of the defender's state space is 28. We assume the defender starts out and remains in state $s^D = \{s_1^D, s_2^D, s_3^D\}$, where for all i, $s_i^D \in \{F, M, L\}$ until an attack occurs. Once an attack occurs, the defender immediately makes transition to the absorbing state Att. We have assumed that the higher the production rate, the higher the level of productivity, and the less time and resources available for mitigating risk, e.g., stopping production to test for toxins.

The attacker's state space S^A The attacker can be in the following states, $S^A = \{TF, IA, AM, AT, AA_i, T_i, i \in \{1, 2, 3\}\}$, where state TF is the attack team formation state; IA is the toxin ingredients acquisition state; AM is the toxin manufacturing state; AT is the toxin transportation state; AA_i is the state where the attacker team is armed for an attack at target i. We assume that the attacker is armed for attack (in AA_i) once the toxin has been manufactured and delivered to the target i. State T_i is the attacked state where an attack occurs at target i.

The state space of the system (S) is the Cartesian product of the defender's state space (S^D) and the attacker's state space (S^A); i.e., $S = S^D \times S^A$. The number of elements in the state space of the system is 280. At decision epoch t, let $s^D(t)$ and $s^A(t)$ be the defender's and attacker's states, respectively, and $s(t) = \{s^D(t), s^A(t)\}$.

4.3.2.2 Action Spaces

The defender's action space A^D The defender can either remain in its current state or move to any other state; hence, the defender is able to choose any predefined productivity level and risk mitigation level for each possible target of the

system. Thus, the cardinality of the action space at each state $(s_1^D, s_2^D, s_3^D), s_i^D \in \{F, M, L\}, i = 1, 2, 3$ is 28.

The attacker's action space A^A We assume that an attack state T_i can only be entered from state AA_i for all i. The usual progression of states is therefore: TF to IA to AM to AT to AA_i. At each decision epoch, the attacker can choose to stay in the current state, advance forward to the next state, or go back to a prior state. For each action at any state except attacked states, there is a probability that the attacker may return to the initial state TF caused by an operational error, interdiction of the defender, etc.

At decision epoch t, let $a^D(t)$ and $a^A(t)$ be the defender's and attacker's actions, respectively, and $a(t) = \{a^D(t), a^A(t)\}$. At each decision epoch, the number of elements in the action space for the system is 84.

4.3.2.3 Observation Spaces

At decision epoch t, let $z^D(t)$ be the possibly inaccurate observation the defender receives of the attacker's state. Similarly, let $z^A(t)$ be the possibly inaccurate observation the attacker receives of the defender's state. Let $z(t) = \{z^D(t), z^A(t)\}$.

The defender's observation space Z^D We assume the observation space of the defender $Z^D = S^A$. Let the matrix of observation probabilities for the defender be $Q^D = \{P(z^D(t)|s^A(t))\}$. We assume that the defender knows that targets i has been attacked at the moment that target i is attacked. Similar to the definition of observation accuracy in [29], we define $Q^D = Q^D(\epsilon^D)$, where $\epsilon^D \geq 0$ is a measure of

defender's observation accuracy (Ortiz, Erera and White, 2013), and

$$Q^D(z^D|s^A) = \begin{cases} 1-\epsilon^D & \text{if } z^D = s^A, s^A \notin \{T_1, T_2, T_3\} \\ \sigma_{s^A z^D}\epsilon^D & \text{if} z^D \neq s^A, s^A \notin \{T_1, T_2, T_3\} \\ 1 & \text{if } z^D = s^A, s^A \in \{T_1, T_2, T_3\} \\ 0 & \text{others} \end{cases}$$

where $\sigma_{s^A z^D} \geq 0, \sigma_{s^A s^A} = 0$ and $\sum_{z^D} \sigma_{s^A z^D} = 1$ for all s^A. Defender's observation quality $Q^D(\epsilon_1^D)$ is considered more accurate than that of $Q^D(\epsilon_2^D)$ if $\epsilon_1^D \leq \epsilon_2^D$. The defender has perfect observation about the attacker if $Q^D = I$, where I is the entity matrix (i.e., $\epsilon^D = 0$).

The attacker's observation space Z^A We assume the observation space of the attacker $Z^A = S^D$. Let the matrix of observation probabilities for the attacker be $Q^A = \{P(z^A(t)|s^D(t))\}$. Similar to Q^D, we define $Q^A = Q^A(\epsilon^A)$, where $\epsilon^A \geq 0$ is a measure of attacker's observation accuracy (Ortiz, Erera and White, 2013), and

$$Q^A(z^A|s^D) = \begin{cases} 1-\epsilon^A & \text{if } z^A = s^D \\ \sigma_{s^D z^A}\epsilon^A & \text{if} z^A \neq s^D \end{cases}$$

where $\sigma_{s^D z^A} \geq 0, \sigma_{s^D s^D} = 0$ and $\sum_{z^A} \sigma_{s^D z^A} = 1$ for all s^D. Attacker's observation quality $Q^A(\epsilon_1^A)$ is considered more accurate than that of $Q^A(\epsilon_2^A)$ if $\epsilon_1^A \leq \epsilon_2^A$. The attacker has perfect observation about the defender if $Q^A = I$, where I is the entity matrix (i.e., $\epsilon^A = 0$).

4.3.2.4 System dynamics

The dynamics of the defender is presented in Figure 4.4, and the dynamics of the attacker is presented in Figure 4.5. The dynamics of the system $P(s(t+1), z(t+1)|s(t), a(t))$ is the Kronecker product of the defender's dynamics and the attacker's

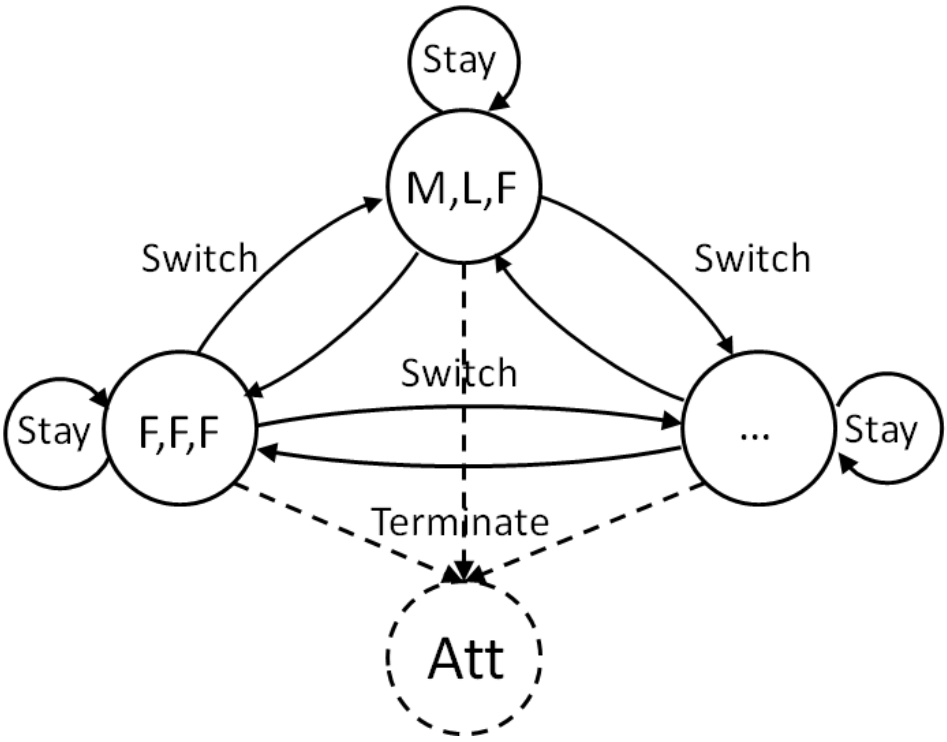


Figure 4.4: Dynamics of the defender

dynamics. Note, $P(s(t+1)|s(t),a(t)) = \sum_{z(t+1)} P(s(t+1),z(t+1)|s(t),a(t))$ and $P(s(t+1),z(t+1)|s(t),a(t)) = Q^D(z^D(t+1)|s^A(t+1))Q^A(z^A(t+1)|s^D(t+1))P(s(t+1)|s(t),a(t))$.

4.3.2.5 Information patterns and policies

The information pattern for agent $k \in \{A, D\}$ describes what agent k knows at each decision epoch. Let the information pattern at time t of finite length τ for agent k be $I^k(t,\tau) = \{s^k(t), \cdots, s^k(t-\tau+1), z^k(t), \cdots, z^k(t-\tau+1), a^k(t-1), \cdots, a^k(t-\tau)\}$, hence, $I^k(t,\tau) = \{s^k(t), z^k(t), a^k(t-1), I^k(t-1,\tau-1)\}$. Let $I^k(0) = \{I^k(0,\tau), y^k(0)\}$, where $I^k(0,\tau)$ is given. $I^k(t) = \{s^k(t), ..., s^k(1), z^k(t), ..., z^k(1), a^k(t-1), ..., a^k(0), I^k(0)\}$ for $t \geq 1$ and $y^k(t)$ is the stochastic array $\{P(I^l(t,\tau)|I^k(t))\}$, $l \neq k$, where $y^k(t)$ is a "belief" array that indicates what agent k can infer about the other agent's information pattern, i.e., $I^l(t,\tau), l \neq k, l, k \in \{A, D\}$. We assume agents make decisions

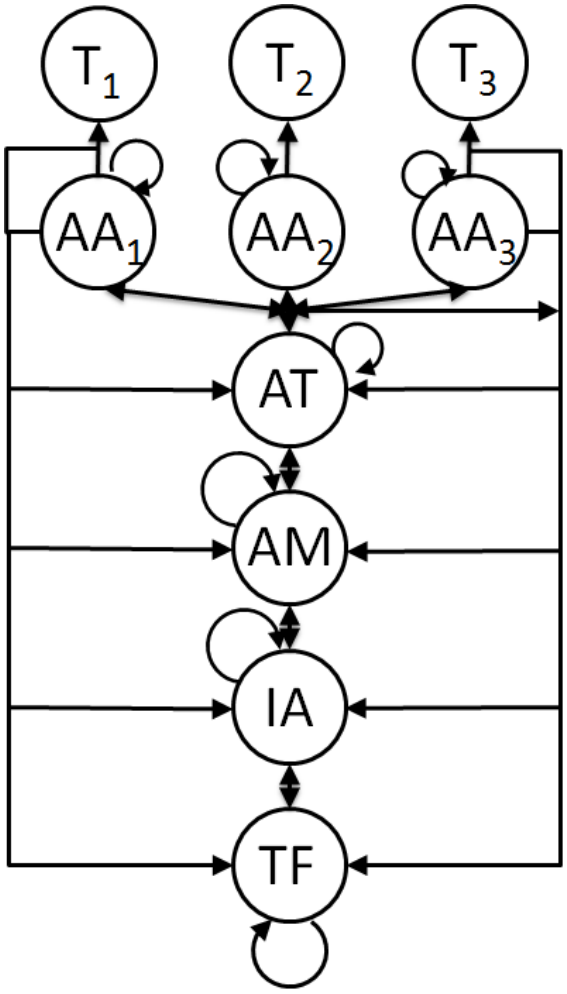


Figure 4.5: Dynamics of the attacker

on the basis of $I^k(t,\tau)$. Hence, a policy π^k for agent $k \in \{A, D\}$ is a mapping from $\{I^k(t,\tau)\}$ to its set of available actions A^k. We focus on stationary policies without loss of generality. Let Π^k be the policy space of agent $k, k \in \{A, D\}$.

4.3.2.6 Rewards, criterion, and objective

Let the single-period reward for the defender and the attacker be $r^D(s,a)$ and $r^A(s,a)$, respectively.

The defender's reward $r^D(s,a)$ The defender has two objectives: $r_1^D(s,a)$ is the productivity measure and $r_2^D(s,a)$ is the risk measure. For the productivity

measure, let $r_1^D(s,a) = \sum_j P(j|s^D, a^D)$[productivity of state j], which we note is independent of the attacker state and action. For the vulnerability measure, let $r_2^D(s,a) = -\rho r^A(s,a)$ where $\rho > 0$ is the coefficient that reflects the fact that the consequence evaluation from the defender's perspective may be different from that of the attacker's.

The attacker's reward $r^A(s,a)$ The attacker has a single objective, and we assume that $r^A(s,a)$ is non-zero only if the attacker attacks the system and then move to $T_i, i \in 1,2,3$. If the attack is successful, then the reward is calculated by the consequence assessment model $cons(m,i)$. We assume that the mass of the toxin inserted into the system (m) depends on the defender's state at the time of the attack and that for each target, $P(m|F) \geq P(m|M) \geq P(m|L)$. A failed attack may result in a penalty cost (c_p) to the attacker.

The criterion used by both agents for all objectives is the expected infinite horizon total discounted reward criterion $v^k(\pi^D, \pi^A)(I^k(0)) = E\{\sum_t \beta^t r^k(s(t), a(t))|I^k(0)\}$ where β is the discount factor such that $0 \leq \beta < 1$. The defender's objectives are to maximize long-run expected total discounted productivity of the food production facility and to minimize the long-run expected total discounted consequence of an attack. The attacker's objective is to maximize the long-run expected total discounted consequence generated by inserting toxin into the food production facility.

4.3.3 Solution procedure

We recall that the defender selects $a^D(t)$ at decision epoch t knowing $I^D(t,\tau)$. Proposition 2.1 in Chapter 2 shows that if the attacker knows the defender's policy, then:

1. The attacker can base selection of an optimal policy at decision epoch t on $s^A(t)$ and the array $\{P(I^D(t,\tau)|I^A(t))\}$, rather than on $I^A(t)$.

2. The optimal value of the attacker's criterion depends on $I^A(t)$ only through $(s^A(t), \{P(I^D(t,\tau)|I^A(t))\})$ and this value is convex and piecewise linear in terms dependent on $s^A(t)$ and $\{P(I^D(t,\tau)|I^A(t))\}$.

We remark that these results are due in part to the fact that the POMG, given a finite-memory policy for the defender, can be transformed into a specially structured partially observed Markov decision process (POMDP) for the attacker (see Chapter 2 for details).

The computational implications of Proposition 2.1 in Chapter 2 are of fundamental importance. Proposition 2.1 guarantees that the determination of an optimal policy and the resulting optimal criterion value for the attacker have finite representation and hence are potentially computable, thus justifying the finite memory assumption for the defender's policy. We base the attacker's action selection at decision epoch t on $\{s^A(t), \{P(I^D(t,\tau)|I^A(t))\}\}$, rather than on $I^A(t)$, because: the number of elements in $(s^A(t), \{P(I^D(t,\tau)|I^A(t))\})$ depends on the fixed constant τ rather than on t, the number of elements in $I^A(t)$ grows linearly in t, and t is unbounded over the infinite planning horizon of the attacker's criterion.

We observe that since the value function of the attacker v^A can now depend on $(s^A(t), \{P(I^D(t,\tau)|I^A(t))\})$, the domain space of v^A is uncountable. However, since v^A is convex and piecewise linear, v^A has the finite representation presented in Proposition 2.1 in Chapter 2. Thus, we are able to compute an optimal policy for the attacker, given a finite-memory policy for the defender (given restrictions on the dimensions of the state, action, and observation spaces).

The specially structured POMDP produces an optimal policy for the attacker that has perfect memory. Proposition 2.2 in Chapter 2 requires that both the defender's

and attacker's policies be finite-memory policies in order for the defender's value function to have finite representation. Thus, Proposition 2.2 in Chapter 2 justifies approximating the attacker's policy that results from solving the POMDP with a finite-memory policy in order to insure potential computability.

We remark that the approach taken in Chapter 2 for determining policies for the defender and attacker satisfies two equilibrium conditions presented in Chapter 2, one for each agent. These equilibrium conditions guarantee that neither the defender nor the attacker can improve its performance by deviating from the equilibrium conditions. Treating v^D as the fitness measure for each defender's policy, a multi-objective genetic algorithm (MOGA) is used in Chapter 2 to determine the defender and attacker policy pairs that satisfies the equilibrium conditions.

Algorithm 1 presents high-level pseudo code for risk assessment determination. Population size and the number of iterations (design parameters) are initialized in line 1-2. Next, the population is initialized by filling it with randomly generated defender policies, and the best defender policies are set to empty. Lines 5 - 17 describe the steps taken by the multi-objective genetic algorithm. For each defender policy $\rho^D \in \Pi^D$, the attacker's best response policy $\pi^A = \pi^*(\rho^D)$ can be obtained by solving a special structured POMDP (Line 7, using Proposition 2.1 in Chapter 2). The defender's value function $v^D(\rho^D, \pi^*(\rho^D))$ can be determined, given a defender policy and its attacker's best response policy (Line 8, using Proposition 2.2 in Chapter 2). The population of defender policies is sorted according to rankings and crowding distances on the basis of v^D (Line 10). The next generation of defender policies is determined from the current generation of defender policies, the fitness measures of each current generation defender policy, a crossover operator, and a mutation operator (Line 15 - 16). This procedure is repeated until the maximum number of iterations N is

achieved, where N is set so that no significant improvement can be obtained (Line 5). The best defender policies are obtained from the pareto frontier of the last population (Line 12 - 14). Detailed discussion of this algorithm can be found in Chapter 2.

Algorithm 1 Risk Assessment Algorithm

```
1: M ← PopulationSize
2: N ← MaxNumberOfIterations
3: population ← initPopulation()
4: BestPolicy ← ∅
5: while n < N do
6:     for i = 1 to M do
7:         Response← POMDPSolver(DefenderPolicy[i])
8:         Fitness[i]←Evaluation(DefenderPolicy[i],Response)
9:     end for
10:    population ← NonDominatedSorting(population)
11:    n = n + 1
12:    if n == N then
13:        BestPolicy←ParetoFrontier(population)
14:    end if
15:    population ← Crossover(population)
16:    population ← Mutation(population)
17: end while
```

4.3.4 Special cases:

Determining $v^D(\rho^D, \pi^*(\rho^D))(I^D(0))$ for any given $\rho^D \in \Pi^D$ as the fitness measure is a crucial step in order to determine the most preferred defender policy in Π^D (Lines 7 - 8). Since different information structures may affect the performances of the defender and the attacker, we now investigate how the defender's performance $v^D(\rho^D, \pi^*(\rho^D))(I^D(0))$ changes over four different scenarios: completely observed case (CO), completely observed defender-partially observed attacker (CDPA), partially observed defender-completely observed attacker (PDCA), and partially observed case (PO).

4.3.4.1 Completely observed case (CO)

We assume that each agent can completely observe the state of the other agent, i.e., $Q^D = I$ and $Q^A = I$. For any defender policy $\rho^D \in \Pi^D$, the resulting POMDP becomes a MDP with 280 states and 3 actions for each state, which can be solved efficiently by standard MDP solution procedures.

4.3.4.2 Completely observed defender-partially observed attacker (CDPA)

We assume that the attacker's state can be completely observed by the defender; however, the defender's state can only be partially observed by the attacker, i.e., $Q^A \neq I$ and $Q^D = I$. For any defender policy $\rho^D \in \Pi^D$, the resulting POMDP involves 280 underlying states, 28 observations, and 3 actions for each state.

4.3.4.3 Partially observed defender-completely observed attacker (PDCA)

We assume that the defender's state is completely observed by the attacker; however, the attacker's state can only be partially observed by the defender, i.e., $Q^A = I$ and $Q^D \neq I$. This special case models the situation where the attacker knows more about the defender than that the defender knows about the attacker, e.g., the defender is a government agency with open records. For any defender policy $\rho^D \in \Pi^D$, the resulting POMDP involves 2800 underlying states and 1 observation and 3 actions for each state.

4.3.4.4 Partially observed case (PO)

We assume that the state of an agent cannot be accurately observed by the other agent, i.e., $Q^A \neq I$ and $Q^D \neq I$. For any defender policy $\rho^D \in \Pi^D$, the resulting POMDP involves 2800 underlying states and 28 observations and 3 actions for each

state.

4.4 Numerical results

Each defender policy is encoded into an array of probability mass vectors. Each row of this array is a probability mass vector over all possible defender actions for every possible $I^D(t, \tau)$. Let $\tau = 1$ for computational simplicity. We assume all defender policies are deterministic since deterministic policies are easy to implement and understand. The cardinality of the defender's policy space is 280^{28} for all cases.

4.4.1 Runtime Results

For the multi-objective genetic algorithm, we assume a population size of 80, one point crossover probability equal to 0.20, and a mutation rate of 7%. We let $N = 50$ (beyond which no significant improvement was determined in our numerical analysis.) The solution procedure in Lin, Bean and White (2004) was implemented to solve the specially structured POMDP for the attacker. We let the discount factor $\beta = 0.45$ to ensure fast convergence. Parallel computing (openMP) was used on a 3.80GHz quad-core (8 threads) CPU. The runtime results are summarized in Table 4.1. We remark that the dynamic programming algorithm presented in Hansen, Bernstein and Zilberstein (2004) can only solve a partially observed stochastic game with 4 states and 2 actions and 2 observations per agent before it runs out of memory.

Table 4.1: Runtime Results

	CO	CDPA	PDCA	PO
underlying state of POMDP	280	280	2800	2800
# of observations of POMDP	1	28	1	28
# of actions (for each state)	3	3	3	3
average runtime for POMDP(sec)	1.2	90.7	55.1	540.5
total runtime (hours)	0.34	24.4	17.3	149.2

4.4.2 Base Model Results

Figure 4.6 shows the improvement of the Pareto frontier from generation to generation for the partially observed case. We terminate the algorithm at the generation 50 beyond which no significant improvement has been found.

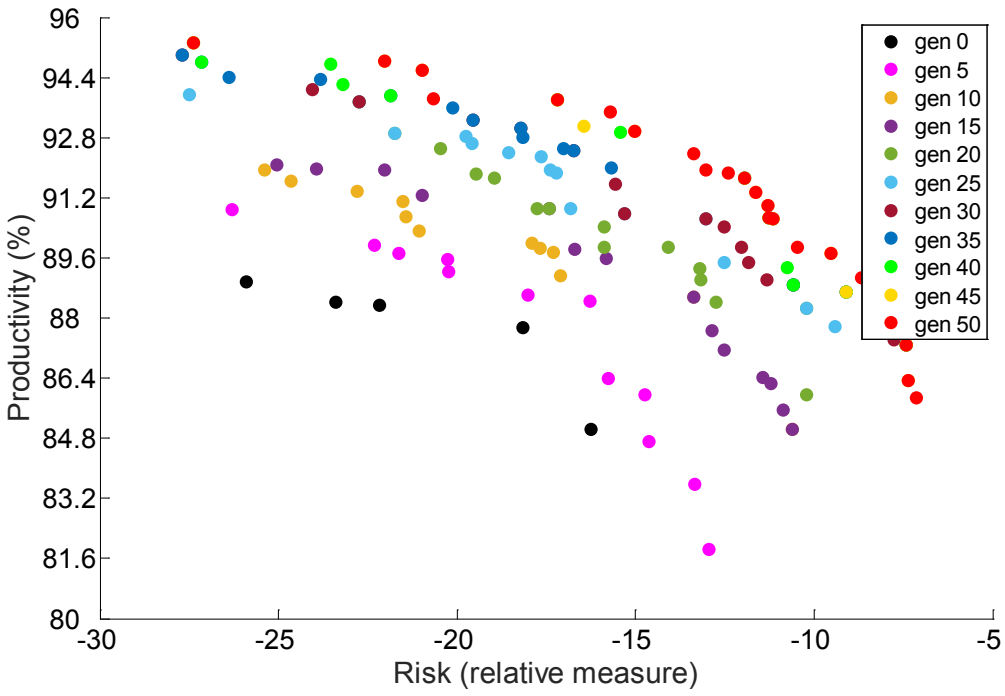


Figure 4.6: Numerical result for partially observed case

Table 4.2 lists the fitness measures of each policy on the Pareto frontier for generation 50, as displayed in Figure 4.6. All of these policies select the action "stopping production and clean the system" when an attack is observed. Policy π_1 is the "least risk, least productivity" policy, where the defender always selects state LLL. The other non-dominated policies $(\pi_2, ...\pi_{23})$ consider the tradeoffs between expected long-run total productivity and expected long-run risk.

We also compare the non-dominated policies with two baseline policies in Table 4.2. Baseline policy b_1 assumes that the defender always select state FFF, presumably with the intent of maximizing productivity. The result shows that always selecting FFF can significantly increase the vulnerability to an attack, which in turn reduces the total long run productivity of the system and hence is dominated. Policies on the Pareto frontier also dominate the baseline policy b_2, which selects action randomly.

4.4.3 Value of Information

We present a comparison of the Pareto frontiers for the four different scenarios in Figure 4.7. The defender's performance depends on the relative observation accuracy about the other agent. The defender's performance in the CDPA case, where the defender can completely observe the attacker but is only partially observed by the attacker, is better than the other three cases. For the PDCA case, where the attacker has accurate information about the defender's state but the defender has inaccurate observations of the attacker, the defender's performance is worst in all the cases. The performance of the defender for the completely observed case (CO) and the partially observed case (PO) is bounded by the defender's performance in the CDPA case (from above) and the defender's performance in the PDCA case (from below). In our example, $\epsilon^D = 0.25 < \epsilon^A = 0.3$; hence, the defender's observation matrix Q^D is

Table 4.2: Non-dominated Policy in partially observed case

Policy	Productivity (%)	Risk
π_1	85.90	-7.15
π_2	86.36	-7.38
π_3	87.30	-7.44
π_4	88.26	-7.67
π_5	88.31	-7.91
π_6	89.09	-8.69
π_7	89.73	-9.53
π_8	89.91	-10.49
π_9	90.64	-11.16
π_{10}	90.68	-11.28
π_{11}	91.00	-11.30
π_{12}	91.36	-11.65
π_{13}	91.72	-11.96
π_{14}	91.88	-12.42
π_{15}	91.96	-13.05
π_{16}	92.37	-13.38
π_{17}	92.99	-15.05
π_{18}	93.50	-15.73
π_{19}	93.81	-17.19
π_{20}	93.84	-20.66
π_{21}	94.62	-20.98
π_{22}	94.85	-22.03
π_{23}	95.33	-27.38
b_1	71.79	-46.14
b_2	87.92	-18.51

slightly better than the attacker's observation matrix Q^A, according to the definition of observation accuracy presented in (Ortiz, Erera and White, 2013). The resulting defender's performance is similar for the completely observed case.

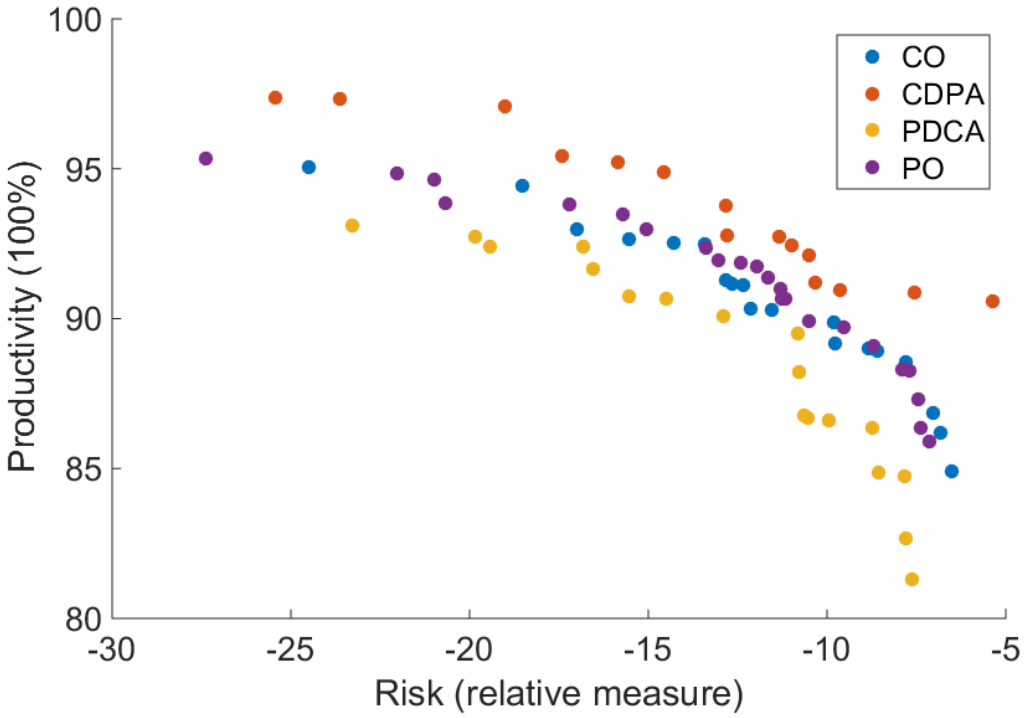


Figure 4.7: Pareto frontiers comparison for four special cases

We now analyze how information asymmetry can affect the defender policy and hence the defender's performance. We select the policy generating the largest measure of productivity from the non-dominated policy set for each of the four special cases. Clearly, the system is at risk when the defender's (possibly inaccurate) observation $z^D \in \{AA_1, AA_2, AA_3\}$. Given such an observation, we would anticipate that the defender would immediately try to make transition into or remain in either state M or L at target i. A defender policy can be viewed as a set of IF-THEN statements of the form: IF the current data available to the defender is $I^D(t, \tau)$, THEN the

defender selects action a^D. For $X \subset \{F, M, L\}$, let $N_i(X)$ be the number of IF-THEN statements that satisfy:

IF $s_i^D(t) \in \{F, M, L\}$ and $z^D(t) = AA_i$, THEN select action $\bar{a}^D$ such that

$$P(s_i^D(t+1) \in X | s_i^D(t), \bar{a}^D) \geq P(s_i^D(t+1) \in X | s_i^D(t), a^D), \forall a^D.$$

The ratios of $N_i(L)/N_i(F, M, L)$ are given in Table 4.3(a), and the ratios of $N_i(L, M)/N_i(F, M, L)$ are given in Table 4.3(b), for all $i \in \{1, 2, 3\}$.

For all $AA_i, i \in \{1, 2, 3\}$, the ratios of $N_i(L)/N_i(F, M, L)$ in PDCA are higher than the ratios of $N_i(L)/N_i(F, M, L)$ in CDPA. The observation AA_i is accurate in CDPA but inaccurate in PDCA. In order to mitigate the risk of an attack, the defender is more likely to make transition into or remain in state L to overcome the additional uncertainty from the inaccurate observation. The ratios of $N_i(L)/N_i(F, M, L)$ in PDCA are also higher than the ratios of $N_i(L)/N_i(F, M, L)$ in PO because: the attacker also receives inaccurate observations about the defender's state in PO, imposing an additional challenge to attack the system, which may reduce the likelihood of an attack. Consequently, the defender may lower the ratio of $N_i(L)/N_i(F, M, L)$ in PO. It is interesting that the ratios of $N_i(L)/N_i(F, M, L)$ in CDPA are higher than the ratios of $N_i(L)/N_i(F, M, L)$ in PO, $\forall i \in \{1, 2, 3\}$. A potential reason is that in CDPA, only the attacker has the uncertainty about the observation and the attacker knows this fact. Hence, the attacker can utilize this fact and affect the performance of the defender. On the contrary, both the defender and the attacker have the uncertainties from the inaccurate observations in PO. And the uncertainties from both sides can cancel out when selecting action.

The result that the ratios of $N_i(L)/N_i(F, M, L)$ in CO are higher than the ratios of $N_i(L)/N_i(F, M, L)$ in CDPA, $i \in \{1, 2, 3\}$ in Table 4.3(a) shows that it is beneficial for the defender to introduce the uncertainty of observation to the attacker, which

Table 4.3: Defender policy in four cases when $z^D \in \{AA_1, AA_2, AA_3\}, s^D \neq Att$

(a) the ratios of $N_i(L)/N_i(F, M, L)$ when observation $z^D = AA_i, i \in \{1, 2, 3\}$

$z^D(t)$	CO	CDPA	PDCA	PO
AA_1	58.44%	55.56%	59.26%	44.44 %
AA_2	51.85%	37.04 %	37.04%	33.33%
AA_3	40.74%	29.63%	37.04%	25.93 %

(b) the ratios of $N_i(L, M)/N_i(F, M, L)$ when observation $z^D = AA_i, i \in \{1, 2, 3\}$

$z^D(t)$	CO	CDPA	PDCA	PO
AA_1	70.37%	81.48%	74.07%	74.07%
AA_2	66.67%	70.37%	70.37%	77.78%
AA_3	62.96%	48.15%	55.56%	51.85%

can reduce the percentage of actions of making transition into or remain in state L. Table 4.3(b) shows that $\frac{N_1(L,M)}{N_1(F,M,L)} \geq \frac{N_2(L,M)}{N_2(F,M,L)} \geq \frac{N_3(L,M)}{N_3(F,M,L)}$, which is consistent with the fact in Figure 4.3 that the target 1 is most vulnerable, target 2 is the next, and target 3 is least vulnerable (The mass of toxin we considered is relatively large). The defender is more likely to make transition into or remain in state L or M for more vulnerable targets.

It is reasonable to consider that the system is at relatively low risk when the defender's (possibly inaccurate) observation $z^D \in \{AT, AM, IA, TF\}$, suggesting that an attack is not imminent. In such situations, we would imagine that the defender would prefer state F at each target in order to improve the system productivity. We assume the more targets are in state F, the more productive the system is. If the number of targets in state F is fixed, the more targets are in state M, the more productive the system is as well. Assume the productivity level of state (F, F, F) is 100%. For $Y \subset \{10j\%, j = 1...10\}$, let $M(Y)$ be the number of IF-THEN statements that satisfy:

IF $s^D(t) \neq Att$ and $z^D(t) \notin \{AA_1, AA_2, AA_3\}$, THEN select action $\tilde{a}^D$ such that $P($ Productivity of $s_i^D(t+1) \in Y | s_i^D(t), \tilde{a}^D) \geq P($ Productivity of $s_i^D(t+1) \in Y | s_i^D(t), a^D), \forall a^D$.

The percentages of $M(10k\%)$ over $M(\{10j\%, j = 1...10\})$ are shown in Figure 4.8 for the four special cases, $k = 2, ..., 10$. Comparing CDPA to PDCA shows that the percentage of $M(10k\%), k \geq 6$ in CDPA is greater than the percentage of $M(10k\%), k \geq 6$ in PDCA. Hence, the defender is more willing to making transition into or remain in states with high productivity level ($\geq 60\%$) if its observation about the attacker is accurate, which may improve the measure of productivity. The result that the percentage of $M(10k\%), k \geq 7$ in CDPA is higher than $M(10k\%), k \geq 7$ in PO also indicates the added value of improving the observation quality of the defender about the attacker. Observe that $M(10k\%), k \geq 7$ in PDCA is greater than $M(10k\%), k \geq 7$ in PO. We have shown that when the system is in high risk, the ratios of $N_i(L)/N_i(F, M, L)$ in PDCA are higher than the ratios of $N_i(L)/N_i(F, M, L)$ in PO, $\forall i \in \{1, 2, 3\}$. Accordingly, the defender has to improve the percentage of selecting a more productive state in order to maintain high productivity. The defender's policy balances the measures of productivity and risk.

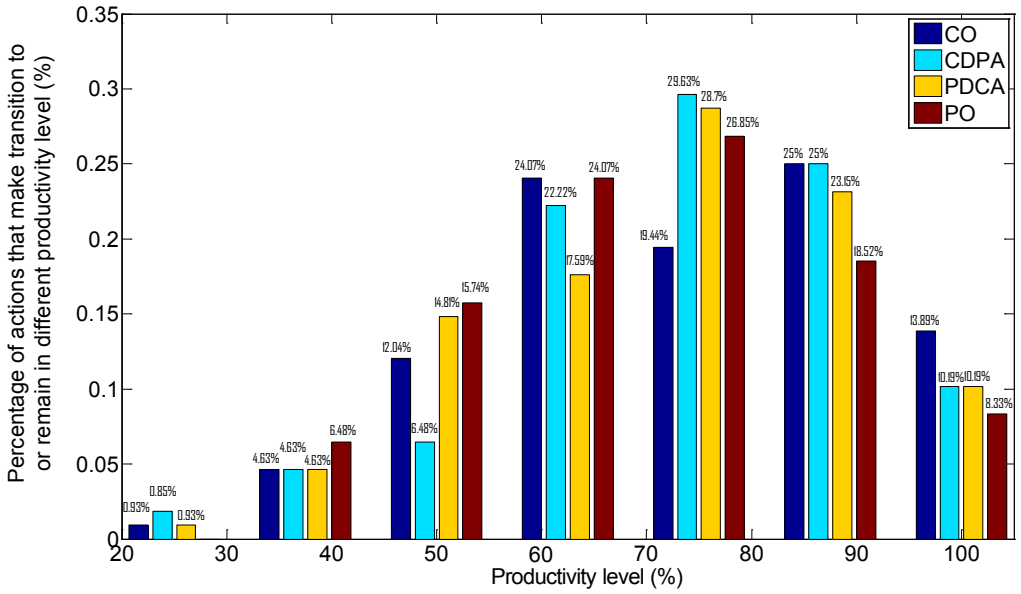


Figure 4.8: Defender policy in four cases when $z^D \in \{AT, AM, IA, TF\}, s^D \neq Att$

The observation quality associated with $Q^D(\epsilon_1^D)$ is considered more accurate than

that of $Q^D(\epsilon_2^D)$ if $\epsilon_1^D \leq \epsilon_2^D$, where ϵ^D can be thought of a measure of observation error. We fix Q^A and further analyze how the defender's performance can progressively change as the observation accuracy improves. Figure 4.9 shows that the performance of the defender tends to improve as ϵ^D decreases, which means improved information accuracy about the attacker can improve the defender's performance. The black stars in Figure 4.9 represent the case where the defender's observations provide no information about the attacker's state; i.e., the $(i,j)^{th}$ element of the probability matrix Q^D is independent of i and hence all rows of Q^D are identical. Thus, the difference between the black stars and any of the non-black shapes (triangle, circle, star, dot, etc.) represents the added value of adaptively and intelligently making use of the information content in the data the defender receives about the attacker's state. The case where only a priori static decision-making is made by the defender is equivalent to the black stars and can be obtained through use of the standard PRA paradigm. A theoretical analysis of value of information for the POMG can be found in Chapter 3.

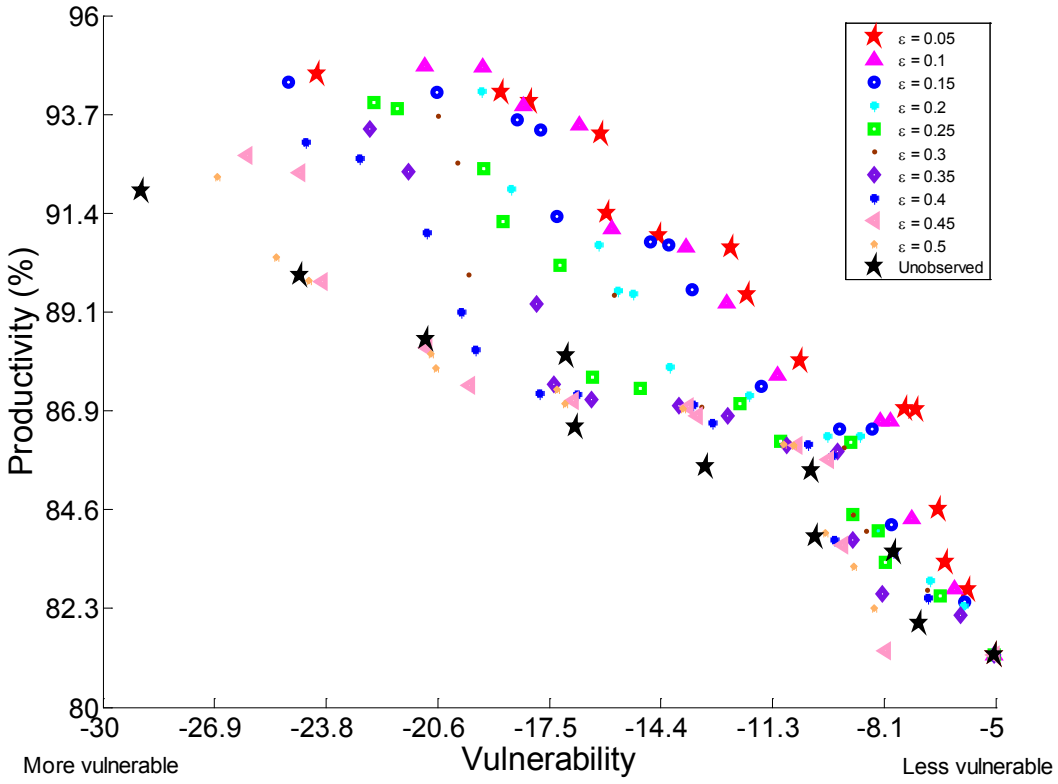


Figure 4.9: Value of information for the defender for various levels of accuracy of the defender's observation matrix $Q^D(\epsilon)$

4.4.4 Dynamic Risk Mitigation

Assume how imminence of an attack increases as the attacker approaches any targets of the system. We define

$$\text{imminence of an attack} = \begin{cases} 0 & \text{if } s^A = TF \\ 1 & \text{if } s^A = IA \\ 2 & \text{if } s^A = AM \\ 3 & \text{if } s^A = AT \\ 4 & \text{if } s^A = AA_i, i \in \{1,2,3\} \\ 5 & \text{if } s^A = T_i, i \in \{1,2,3\} \end{cases}$$

Figure 4.10 shows two sample paths simulated under two defender policies with their corresponding best response attacker policies. The blue line is from the non-dominated defender policy π_2, and the red line is from the baseline policy b_1 where the defender always selects the highest productivity level for all targets. The plot shows how the risk of the system can be mitigated over time as the defender and attacker interact. The attacker progressively moves from TF to AT and then selects its most preferred target. When the attacker is ready to attack a target, we labeled the name of the target the attacker may attack, the defender's state, and defender's action. Attacking Collecting Vat can generate the largest consequence if the risk mitigation is low at every target. The red line shows that if the system is not well protected, the system will be attacked at Collecting Vat quickly (e.g. the second peak), even if there is a probability that the attacker may not be able to attack the system successfully for every attempt (e.g. the first peak).

In contrast, under the non-dominated defender policy π_2, the attacker may shift to different targets while the defender is interacting with the attacker. For example, at period 167, the attacker is ready to attack Collecting Vat. The defender makes transition from (M, F, F) to (L, M, F) to protect the system. Accordingly, the attacker leaves Collecting Vat and moves to Raw Production Tank in the next period. At period 343, the defender makes another transition from (L, M, M) to (M, L, F) since the attacker is ready to attack Raw Product Tank. At period 392, the defender fails to recognize that the attacker is ready to attack Finished Product Tank because of inaccurate information and an attack occurs. The defender is mitigating the risk by dynamically making transitions among various states in response to the attacker's state and action. We observe that during these interactions, the target generating the largest consequence may no longer be attacker's best target to attack, and the attacker may attack a target that is not very vulnerable. Hence, the defender can

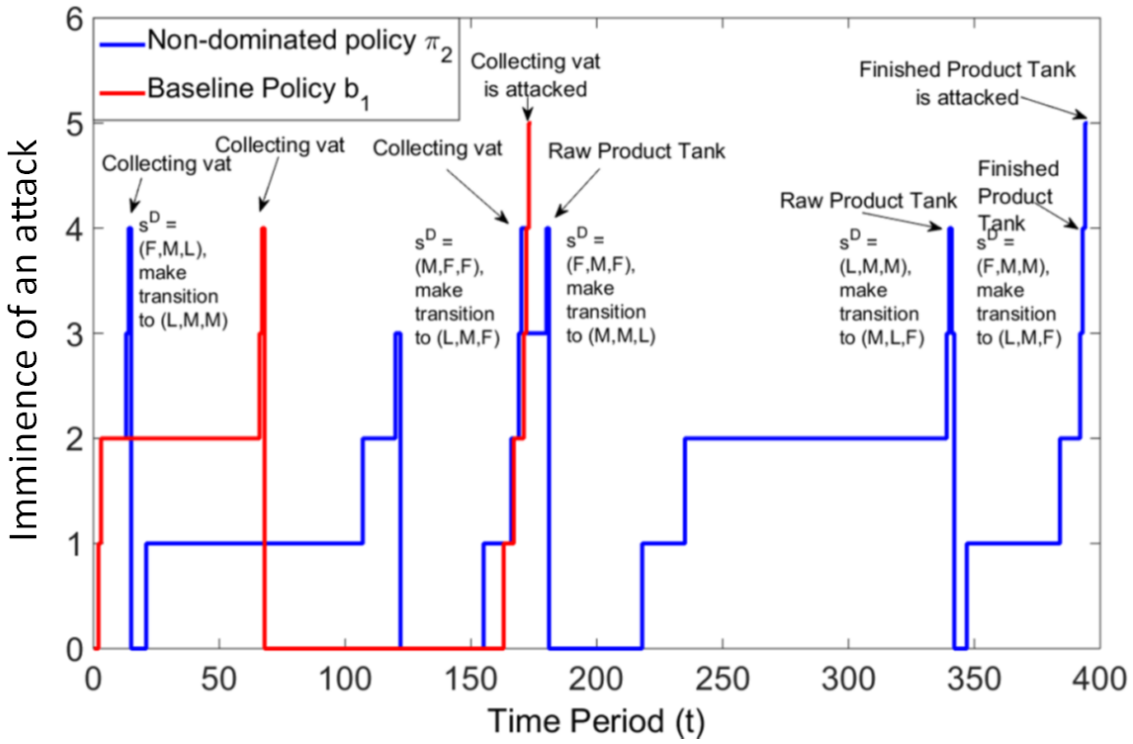


Figure 4.10: Sample paths simulated under two defender's policies with their corresponding best response attacker's policies.

influence the attacker to attack a less vulnerable target. Figure 4.11 shows the percentages of attacks that occur at each target, based on 1000 simulations with policy π_2. Only 2.5% of the attacks occur at Collecting Vat, whereas the attacker will always try to attack Collecting Vat under baseline policy b_1.

Moreover, our model assumes that $P(s^A(t+1) \in \{TF, IA\} | s^A(t), a^A(t)) > 0, \forall s^A(t) \notin \{T_1, T_2, T_3\}$, because for example the toxin may be no longer active or some team members may be not available any more. Figure 4.10 shows that the attacker may have to go back to low-risk states (e.g. TF or IA) while the attacker intends to move to other targets in response to the defender's action. Returning back to low-risk attacker's states extends the time until an attack, which in turn reduces the long-run adversarial risk of the system.

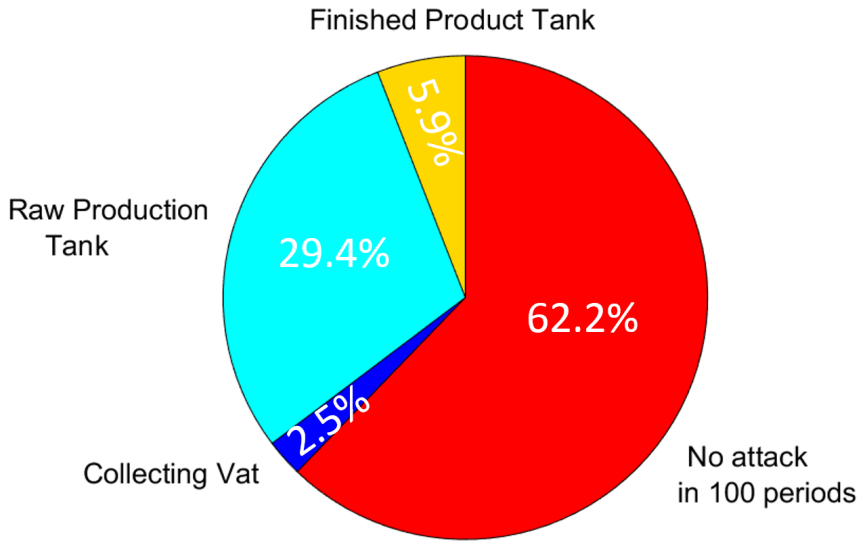


Figure 4.11: The distribution of attacked targets under policy π_2

We show the distribution, as a function of the number of decision epochs, of the time until an attack by a box plot in Figure 4.12, where 1000 sample paths were simulated for each of four policies. The mean, standard derivation (std) and coefficient of variation (CV) of the time until an attack are listed for each policy. Consistent with Table 4.2, the average number of decision epochs until an attack under baseline policy b_1 is smaller than the average number of decision epochs until an attack under the non-dominated policies. Note also that the interval between the 25% quantile and the 75% quantile for policy b_1 is smaller than the intervals for the other policies. The spread (using std as a measure of spread) of the distribution of the number of decision epochs until an attack is larger for policies under which the system is less vulnerable (e.g. π_1 or π_5).

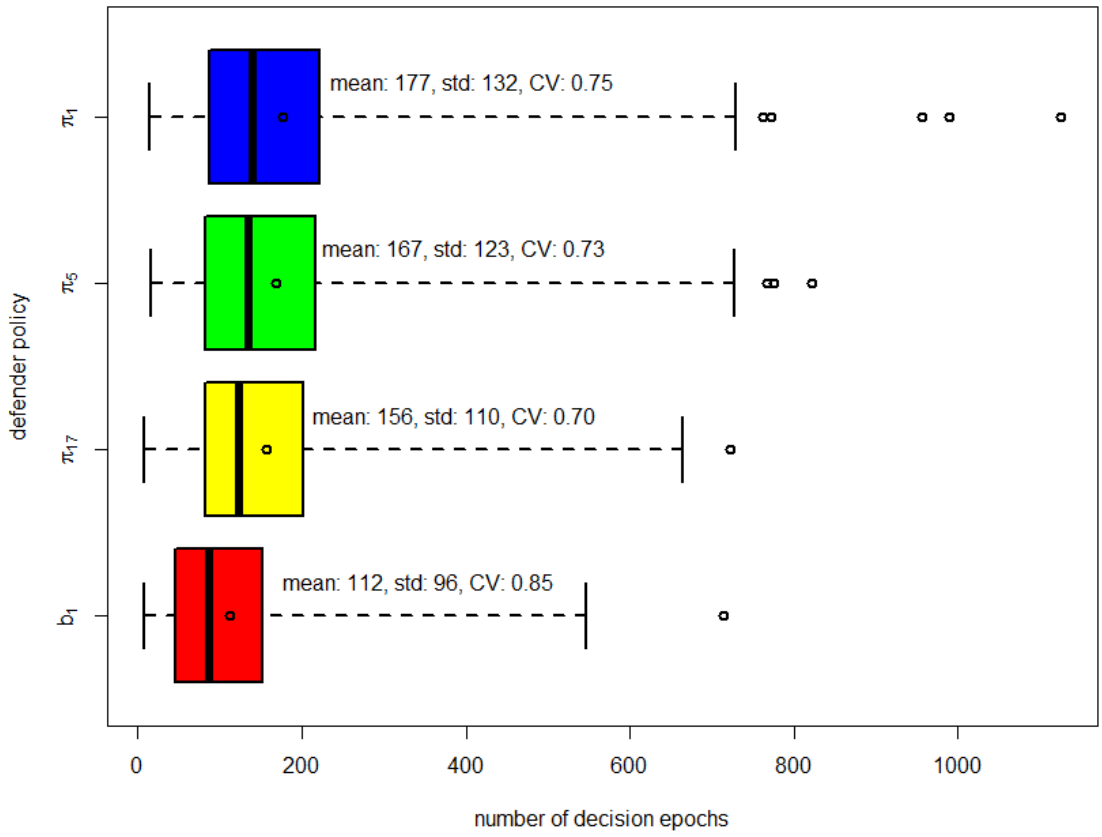


Figure 4.12: Box Plot of the distribution of the time until an attack

4.4.5 Sensitivity Analysis

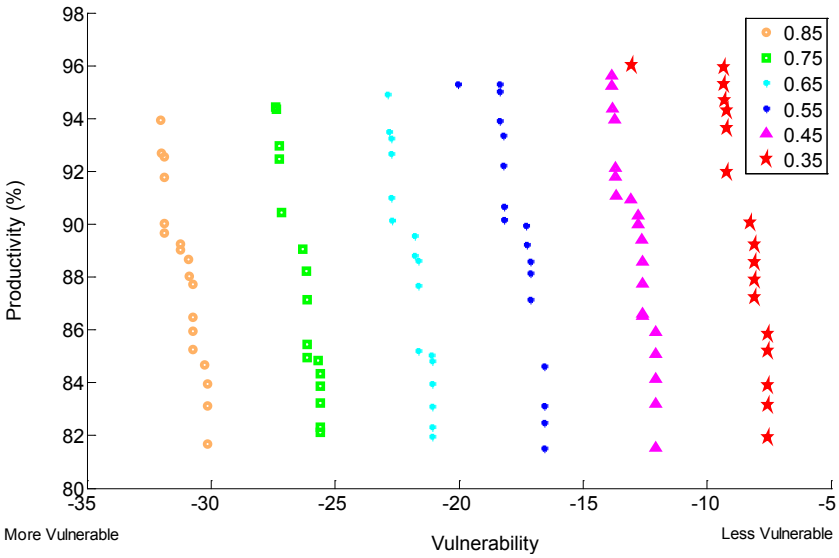
Since some parameters of the risk analysis tool are difficult to estimate, we now examine how changes in critical parameter values can cause changes in defender performance. Transition probabilities are the main unknown parameters in the model since data on which to base estimates of these probabilities are in general unavailable and hence assessment of these probabilities must often rely on expert opinion. Figure 4.13 shows the effect of varying transition probabilities on the Pareto frontier of the defender. Among all of these probabilities, the performance of the defender is particularly sensitive to both the probability of successfully executing an attack and the probability of successfully transporting the toxin to the target. In contrast, changes

in the probabilities of successfully forming a team and acquiring ingredients to make the contaminant cause little change in the performance of the defender. Hence, if the defender can interdict an attack (thereby reducing the probability of successfully transporting the toxin and the probability that an attack will be successful), then investment in interdiction may be a top priority.

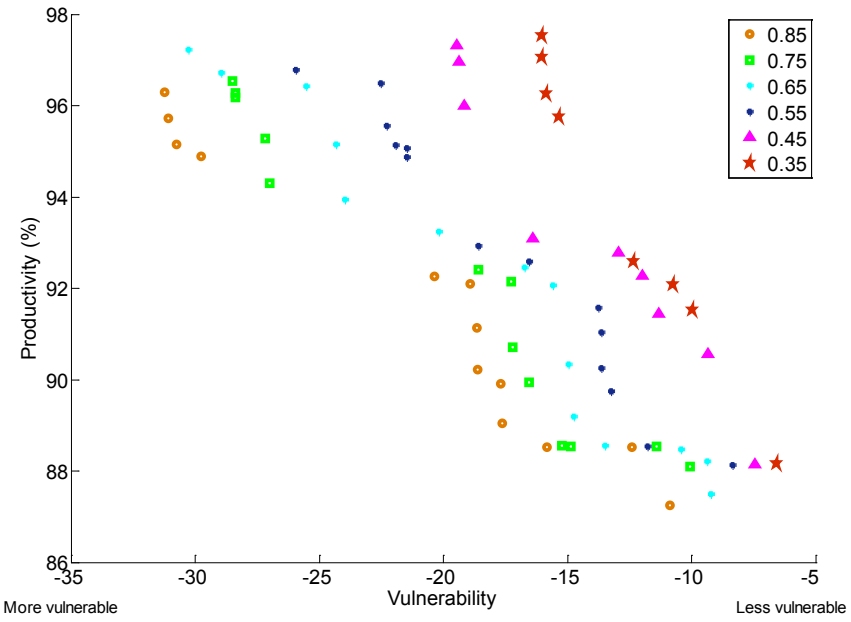
The penalty of an unsuccessful attack on the attacker (c_p) is also difficult to estimate. Intuitively, the higher this penalty, the lower the probability that the attack will want to attack the system. Figure 4.14 supports this intuition and presents how the ratio of c_p to the worst consequence that the system can generate affects the defender's Pareto frontier. Thus, the Pareto frontier moves towards lower productivity and higher vulnerability as this ratio decreases, and hence the higher c_p is, the better it is for the defender. There are two important thresholds: (1) the value of c_p below which the performance of the defender may be significantly decreased and the attacker tends to attack more often because of a low cost of failure for the attacker; (2) the value of c_p above which the attacker is not willing to attack the system because of the high penalty of failure. Hence, the defender should consider investing in raising c_p to at least above the latter ratio in order to protect the facility from attack. However, investment greater than the latter ratio is unnecessary.

4.5 Conclusions

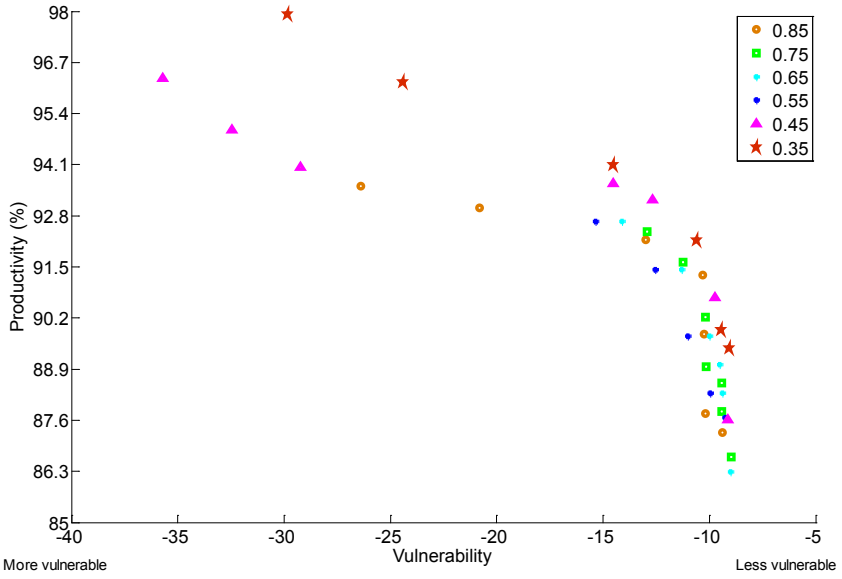
The intent of the risk assessment approach presented in this chapter is to describe the strategic interaction between intelligent and adaptive agents with different objectives over an at most countable number of decision epochs. This approach has two components: (1) a consequence assessment tool and (2) a game theoretic optimization model, where the consequence assessment tool serves as the input to the reward



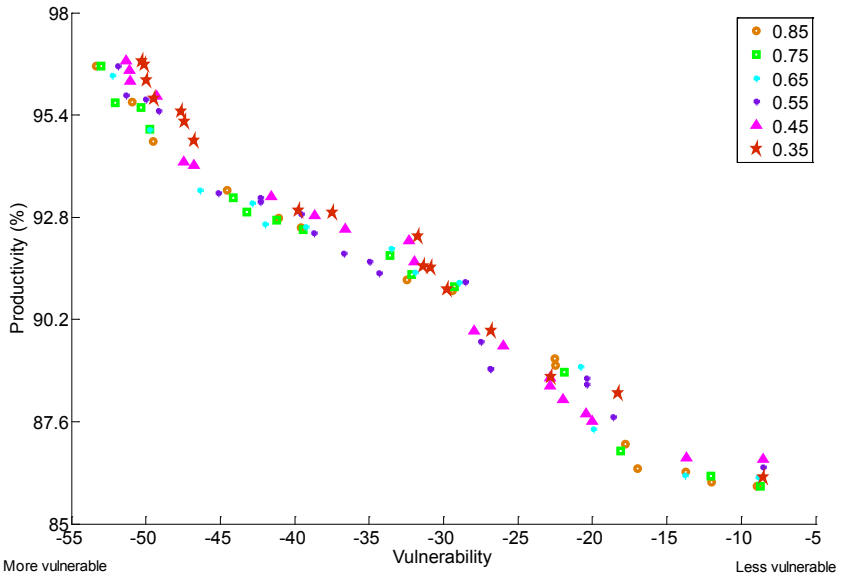
(1) Probability of moving to attacked state T_i from AA_i



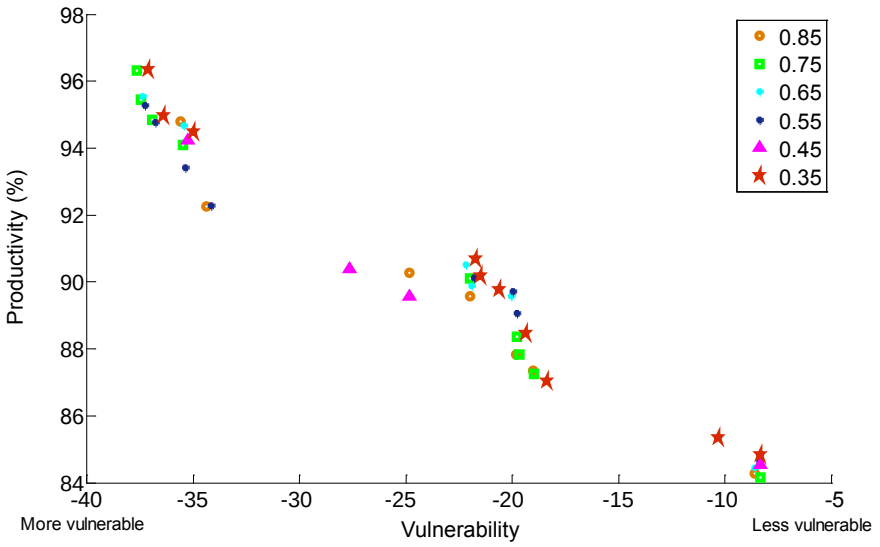
(2) probability of moving to state AA_i from toxin transportation state AT



(3) probability of moving to toxin transportation state AT from manufacturing state AM



(4) probability of moving to manufacturing state AM from ingredient acquisition state IA



(5) probability of moving to ingredient acquisition state IA from team formation state TF

Figure 4.13: Sensitivity analysis for transition probability

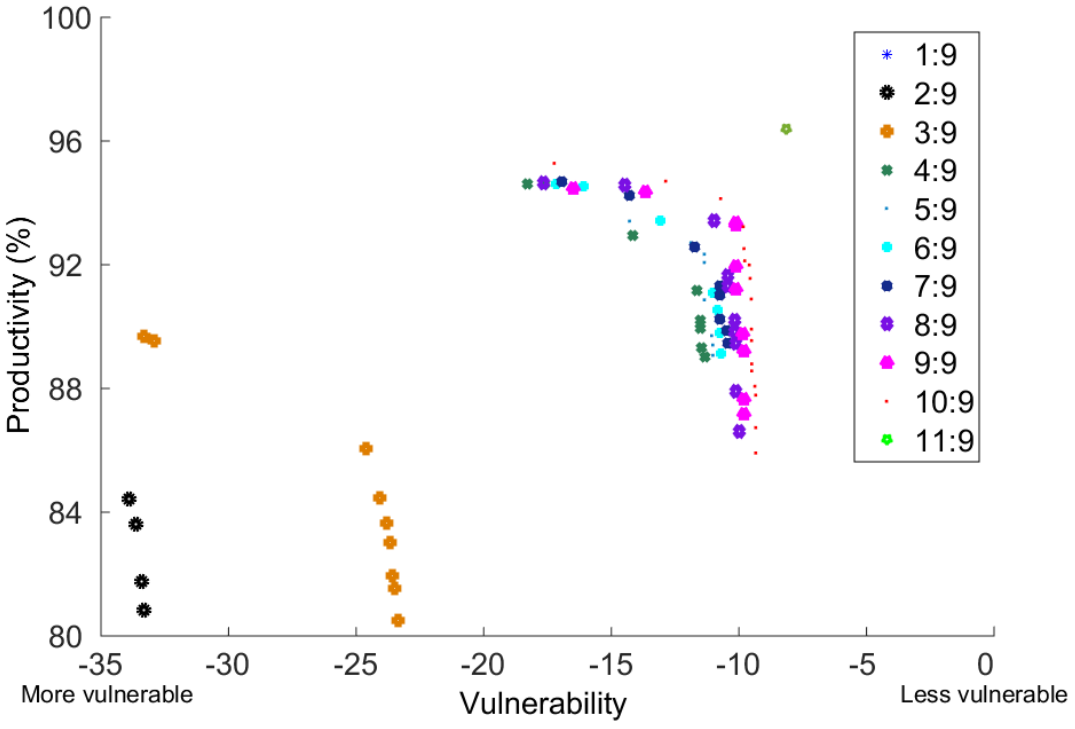


Figure 4.14: Penalty of unsuccessful attack c_p

function in the optimization model. Our risk analysis model can be used to describe the situation where there are two agents, the defender and the attacker, who are both intelligent and adaptive, have different objectives, have a sequence of decisions to make over time, and have different data sets on which to base these decisions. We considered several different variations of the model, where these variations were distinguished by the quality of observation that one agent has of the other agent's current state. In particular, our approach provides decision support to the defender on when and what action should be taken in order to achieve the defender's possibly multiple objectives.

To illustrate our approach, we considered the management of a simplified liquid egg production plant, where: (1) the defender (e.g., the plant manager) has to balance achieving two objectives, maximize plant productivity and minimize the expected consequence of deliberate contamination, and (2) the attacker's objective is to maximize the total number of contaminated packaged units leaving the plant. Our preliminary analysis shows that the system is under greatest risk if the attacker can accurately observe the defender's state and the defender can only inaccurately observe the attacker's state. We showed the impact on risk reduction of reducing the attacker's observation accuracy of the defender. We evaluated the defender's performance, as a function of the defender's observation accuracy of the attacker, indicating the significant value-added that observation accuracy can play in such situations. We performed a sensitivity analysis to better understand what parameter values need careful assessment and what parameters do not.

4.6 References

[1] N. O. Bakir and K. Kardes, "A stochastic game model on overseas cargo container security", In *IEEE International Conference Technologies for Homeland Security (HST)*, pp.110 - 116, New York, 2011.

[2] T. Bedford and R. Cooke, "Probabilistic risk analysis: foundations and methods", *Cambridge University Press*, Cambridge, UK, 2001.

[3] V. M. Bier, N. Haphuriwat, J. Menoyo, R. Zimmerman, and A. M. Culpen, "Optimal resource allocation for defense of targets based on differing measures of attractiveness", *Risk Analysis*, vol.28, no.3, pp.763 - 770, 2008.

[4] V. M. Bier, S. Oliveros, and L. Samuelson "choosing what to protect: Strategic defensive allocation against an unknown attacker", *Journal of Public Economic Theory*, vol.9, no.4, pp.563 - 587, 2007.

[5] E. Bompard, C. Gao, R. Napoli, A. Russo, M. Masera, and A. Stefanini, "Risk assessment of malicious attacks against power systems", *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol.39, no.5, pp.1074 - 1085, 2009.

[6] G. G. Brown and L. A. Cox, "How probabilistic risk assessment can mislead terrorism risk analysis", *Risk Analysis*, vol.31, no.2, pp.196 - 204, 2011.

[7] G. Brown, M. Carlyle, J. Salmeron, and K. Wood, "Defending critical infrastructure", *Interfaces*, vol.36, no.6, pp.530 - 544, 2006.

[8] M. Brown, B. An, C. Kiekintveld, F. Ordonez, M. Tambe, "Multi-objective optimization for security games", In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, vol.2, pp.863 - 870, 2012.

[9] G. G. Brown and L. A. Cox, "How probabilistic risk assessment can mislead terrorism risk analysis", *Risk Analysis*, vol.31, no.2, pp.196 - 204, 2011.

[10] E. Clark and D. Philpott, "CARVER+Shock vulnerability assessment tool", *Government Training Inc.*, 2011.

[11] L. A. Cox, "Some limitations of risk = threat × vulnerability × consequence for risk analysis of terrorist attacks", *Risk Analysis*, vol.28, no.6, pp.1749 - 1761, 2008.

[12] L. A. Cox, "Improving risk-based decision making for terrorism applications", *Risk Analysis*, vol.29, no.3, pp.336 - 341, 2009.

[13] Department of homeland security office of inspector general, *The Department of Homeland Security's Role in Food Defense and Critical Infrastructure Protection*, February 2007.

[14] B. C. Ezell, S. P. Bennett, D. von Winterfeldt, J. Sokolowski and A. J. Collins, "Probabilistic risk analysis and terrorism risk", *Risk Analysis*, vol.30, no.4, pp.575 - 589, 2010.

[15] G. Feichtinger and A. J. Novak, "Terror and counterterror operations: differential games with cyclical Nash solution", *Journal of Optimization Theory and Applications*, vol.139, no.3, pp.541 - 556, 2008.

[16] Food, Drug Administration Center for Food Safety, Joint Institute for Food Safety Applied Nutrition (FDA/CFSAN), Applied Nutrition (JIFSAN), and Risk Sciences International (RSI), "Fda-irisk version 1.0. fda cfsan", *Technical report*, College Park, Maryland. Available at http://irisk.foodrisk.org/., 2012.

[17] B. Golany, E. H. Kaplan, A. Marmur, and U. G. Rothblum, "Nature plays with dice - terrorists do not: allocating resources to counter strategic versus probabilistic risks", *European Journal of Operations Research*, vol.192, no.1, pp.198 - 208, 2009.

[18] E. Hansen, D. Bernstein, and S. Zilberstein, "Dynamic programming for partially observable stochastic game", In *Proceedings of the Nineteenth National Conference on Artificial Intelligence*, pp.709-715, San Jose, California, 2004.

[19] M. Hao, S. Jin and J. Zhuang, "Robustness of optimal defensive resource allocations in the face of less fully rational attacker", In *Proceedings of the 2009 Industrial Engineering Research Conference*, pp. 886 - 891, Norcross, GA, 2009.

[20] K. Hausken and J. Zhuang, "Governments' and terrorists' defense and attack in a t-period game", *Decision Analysis*, vol.8, no.1, pp.46 - 70, 2011.

[21] K. Hausken and J. Zhuang, "The timing and deterrence of terrorist attacks due to exogenous dynamics", *Journal of the Operational Research Society*, vol.63, pp.726 - 735, 2012.

[22] Headquarters Marine Corps, "Operational risk management", *Technical report*, Washington, DC, 2002.

[23] S. Hora, "Eliciting probabilities from experts", *Advances in Decision Analysis, Cambridge University Press*, Cambridge, UK, In Edwards W, Miles R, Jr., von Winterfeldt D. edition, 2007.

[24] G. Levitin and K. Hausken, "Redundancy vs. protection in defending parallel systems against unintentional and intentional impacts", *IEEE Transactions on Reliability*, vol.58, no.4, pp.679 - 690, 2009.

[25] G. Levitin and K. Hausken, "Influence of attacker's target recognition ability on defense strategy in homogeneous parallel systems", *Reliability Engineering and System Safety*, vol.95, pp.565 - 572, 2010.

[26] Z. Lin, J. C. Bean, and C. C. White, "A hybrid genetic/optimization algorithm for finite-horizon, partially observed Markov decision processes", *INFORMS Journal on Computing*, vol. 16, no. 1, pp.27-38, 2004.

[27] W. L. McGill, B. M. Ayyub, and M. Kaminskiy, "Risk analysis for critical asset protection", *Risk Analysis*, vol.27, no.5, pp.1265 - 1281, 2007.

[28] National Center for Food Protection and Defense, "National center for food protection and defense final report 2007 - 2011", *Technical report*.

[29] O. L. Ortiz, A. L. Erera, and C. C. White, "State observation accuracy and finite-memory policy performance", *Operations Research Letters*, vol. 41, pp.477-481, 2013.

[30] R. Powell, "Defending against terrorist attacks with limited resources", *American Political Science Review*, vol.101, no.3, pp.527 - 541, 2007a.

[31] R. Powell, "Allocating defensive resources with private information about vulnerability", *American Political Science Review*, vol.101, no.4, pp.799 - 809, 2007b.

[32] H. Rosoff and D. von Winterfeldt, "A risk and economic analysis of dirty bomb attacks on the ports of los angeles and long beach", *Risk Analysis*, vol.27, no.3, pp.533 - 546, 2007.

[33] T. Sandler and K. Siqueira, "Games and terrorism: Recent developments", *Simulation & Gaming*, vol.40, no.2, pp.164 - 192, 2009.

[34] X. Shan and J. Zhuang, "Cost of equity in homeland security resource allocation in the face of a strategic attacker", *Risk Analysis*, vol.33, no.6, pp.1083 - 1099, 2013.

[35] United States Census Bureau, "US Gazetteer files: 2010, 2000, and 1990", February 2011.

[36] U.S. Department of Health and Human Services, Food and Drug Administration, Center for Food Safety and Applied Nutrition, "Guidance for industry: food producers, processors, and transporters: food security preventive measures guidance", October 2007.

[37] U.S. Food and Drug Administration, "FDA budget requests $4.7 billion to ensure safety of food supply and to modernize medical product safety", *The President's FY 2014 budget for the FDA*, April 2013.

[38] Y. Zhang, "Contributions in supply chain risk assessment and mitigation", *PhD thesis*, School of Industrial and Systems Engineering, Georgia Institute of Technology, 2013.

[39] J. Zhuang, V. M. Bier, and O. Alagoz, "Modeling secrecy and deception in a multiple period attacker-defender signaling game", *European Journal of Operational Research*, vol.203, pp.409 - 418, 2010.

[40] J. Zhuang and V. M. Bier, "Reasons for secrecy and deception in homeland-security resource allocation", *Risk Analysis*, vol.30, no.12, pp.1737 - 1743, 2010.

[41] J. Zhuang and V. M. Bier, "Secrecy and deception at equilibrium with applications to anti-terrorism resource allocation", *Defence and Peace Economics*, vol.22, no.1, pp.43 - 61, 2011.

CHAPTER V

CONCLUSIONS AND FUTURE RESEARCH

This dissertation examines a partially observed Markov game (POMG). The POMG models a sequential decision making situation with multiple intelligent and adaptive decision makers, each of which can choose actions that affect the dynamics of the system and where these actions are selected on the basis of current but possibly inaccurate state observations. The POMG is a new, relatively unexamined combination of the stochastic game and the partially observe Markov decision process (POMDP). This dissertation considers the case where there are two decision makers, a leader and a follower. The leader is allowed to consider multiple objectives in selecting its policy, and the follower considers a single objective in selecting its policy with complete knowledge of and in response to the policy selected by the leader. The decision makers can be cooperative, non-cooperative, or a mixture of both.

Chapter 2 develops a heuristic approach in order to generate a set of non-dominated finite-memory policies from which one of two agents (the leader) can select a most preferred policy to control a dynamic system that is also affected by the control decisions of the other agent (the follower). Each agent's policy assumes that the agent knows its current and recent state values, its recent actions, and the current and recent possibly inaccurate observations of the other agent's state. For each candidate finite-memory leader policy, we assume the follower, fully aware of the leader policy, determines a policy that optimizes the follower's criterion. The leader-follower assumption allows the POMG to be transformed into a specially structured, partially observed Markov decision process that we use to determine the follower's best response policy for a

given leader policy. We then present a value determination procedure to evaluate the performance of the leader for a given leader policy, based on which non-dominated set of leader polices can be selected by existing heuristic approaches (e.g. Multi-objective genetic algorithms).

We analyze how the value of the leader's criterion changes due to changes in the leader's quality of observation of the follower in Chapter 3. We give conditions that insure improved observation quality will improve the leader's value function, assuming that changes in the observation quality do not cause the follower to change its policy. We show that discontinuities in the value of the leader' criterion, as a function of observation quality, can occur when the change of observation quality is significant enough for the follower to change its policy. We present conditions that determine when a discontinuity may occur and conditions that guarantee a discontinuity will not degrade the leader's performance.

This approach is applied in Chapter 4 to quantify the risk of a food production facility to an intelligent and adaptive adversary intent on delivering a chemical or biological toxin. We assume that both the manager (defender) of the food production facility and the perpetrator (attacker) select actions at each of up to a countable number of decision epochs on, in part, the basis of possibly inaccurate information about the other agent. The defender's objectives are to maximize expected facility productivity and to minimize the expected consequence of food contamination. The attacker's objective is to maximize its reward, which combines the long-run expected total discounted consequence of an attack with a penalty if the attack is unsuccessful. We model this problem as a leader-follower, two agent partially observed Markov game, a new model of dynamic risk analysis, and apply a new approach for analyzing risk. We show how the risk can be mitigated as the result of strategic interaction between

two agents over time. We investigate the impact of observation accuracy on facility productivity and risk, thus providing a measure of the value of information, and perform a sensitivity analysis on key parameters.

The POMG developed so far assumes there are a single leader and a single follower. Future research includes extending this model to the situation where there are multiple followers. Each of the followers may have different objectives and they may not share the same information pattern. The followers can communicate with each other and hence collaborate or possibly compete with each other. The objective and information pattern for the leader can also be different from those of the followers. This framework can model many real applications. For example, it can describe the situation where the defender protects the infrastructure against multiple adversaries, or the situation where the manager leads a team to complete a complex task in the business environment, where each member of the team has a different objective function and a different information pattern. It will be interesting to determine the value of improving the ability of the followers to communicate with each other in a collaborative game, as well as the value of disrupting communication between followers in a security application.

Furthermore, leader-follower assumption may not be applicable in many applications. The future research should consider different strategic relationship among agents and analyze how the strategic relationship can affect the performance of each agent. For example, the agents may select policies at the same time or the role of leader and follower may be switched.

In addition, the risk of an intentional attack is just one of many types of risk. Our risk assessment tool can model a variety of types of risk, including risk of intentional

or unintentional contamination for food products, disruptions due to tier 3 or 4 ingredient supplier failure, carrier failure, and natural hazards. The multi-objective characteristic of our model allows us to consider multiple risks simultaneously by treating each risk measure as a separate objective. This method has been used to evaluate the risk of Foot-and-mouth disease to the pork industry.

Appendices

The parameters in Example 3.1 are:

- transition probabilities:

$$P^L(1) = \begin{bmatrix} 0.6229 & 0.3771 \\ 0.7506 & 0.2494 \end{bmatrix}, P^L(2) = \begin{bmatrix} 0.7531 & 0.2469 \\ 0.1761 & 0.8239 \end{bmatrix}$$

$$P^F(1) = \begin{bmatrix} 0.2232 & 0.7768 \\ 0.5131 & 0.4869 \end{bmatrix}, P^F(2) = \begin{bmatrix} 0.9449 & 0.0551 \\ 0.2663 & 0.7337 \end{bmatrix}$$

- reward structure $r^k(s^L, s^F, a^L, a^F), k \in \{L, F\}$:

$$r^F = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{c} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \end{array} \\ \begin{bmatrix} 3.9855 & 1.2631 & 3.8957 & 4.9839 \\ 8.3138 & 8.6463 & 4.9014 & 5.2923 \\ 1.8500 & 7.6665 & 8.8690 & 9.0970 \\ 5.0079 & 5.6435 & 9.0505 & 5.7860 \end{bmatrix} \end{array}$$

(i) example(a):

$$r^L = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{c} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \end{array} \\ \begin{bmatrix} 6.8156 & 8.2357 & 8.9439 & 9.5346 \\ 4.6326 & 1.7501 & 5.1656 & 5.4088 \\ 2.1216 & 1.6357 & 7.0270 & 6.7973 \\ 0.9852 & 6.6599 & 1.5359 & 0.3656 \end{bmatrix} \end{array}$$

(ii) example(b):

$$r^L = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{c} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \end{array} \\ \begin{bmatrix} 7.9466 & 7.5195 & 6.7120 & 3.9076 \\ 5.7739 & 2.2867 & 7.1521 & 8.1614 \\ 4.4004 & 0.6419 & 6.4206 & 3.1743 \\ 2.5761 & 7.6733 & 4.1905 & 8.1454 \end{bmatrix} \end{array}$$

(iii) example(c):

$$r^L = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{c} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \end{array} \\ \begin{bmatrix} 8.0549 & 8.8651 & 9.7868 & 0.5962 \\ 5.7672 & 0.2867 & 7.1269 & 6.8197 \\ 1.8292 & 4.8990 & 5.0047 & 0.4243 \\ 2.3993 & 1.6793 & 4.7109 & 0.7145 \end{bmatrix} \end{array}$$

(iv) example(d):

$$r^L = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{c} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \end{array} \\ \begin{bmatrix} 4.9417 & 8.9092 & 0.3054 & 9.0472 \\ 7.7905 & 3.3416 & 7.4407 & 6.0987 \\ 7.1504 & 6.9875 & 5.0002 & 6.1767 \\ 9.0372 & 1.9781 & 4.7992 & 8.5944 \end{bmatrix} \end{array}$$

(v) example(e):

$$r^L = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{c} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \end{array} \\ \begin{bmatrix} 6.3114 & 9.9685 & 4.3000 & 0.6463 \\ 8.5932 & 5.5354 & 4.9181 & 4.3618 \\ 9.7422 & 5.1546 & 0.7104 & 8.2663 \\ 5.7084 & 3.3068 & 8.8774 & 3.9453 \end{bmatrix} \end{array}$$

(vi) example(f):

$$r^L = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{c} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \end{array} \\ \begin{bmatrix} 0.8348 & 8.9075 & 9.2831 & 8.6271 \\ 6.2596 & 9.8230 & 5.8009 & 4.8430 \\ 6.6094 & 7.6903 & 0.1698 & 8.4486 \\ 7.2975 & 5.8145 & 1.2086 & 2.0941 \end{bmatrix} \end{array}$$

The parameters in Example 3.2 are:

- transition probabilities: $P(s(t+1), z^F(t+1)|s(t), \pi^L, \rho^F) =$

	$1_s^L,1_z^F,1_s^F$	$1_s^L,1_z^F,2_s^F$	$1_s^L,2_z^F,1_s^F$	$1_s^L,2_z^F,2_s^F$	$2_s^L,1_z^F,1_s^F$	$2_s^L,1_z^F,2_s^F$	$2_s^L,2_z^F,1_s^F$	$2_s^L,2_z^F,2_s^F$
$(1^L,1^F)_s,(1^L,1^F)_a$	0.2371	0.1766	0.1741	0.1291	0.1268	0.0990	0.0479	0.0095
$(1^L,1^F)_s,(1^L,2^F)_a$	0.2371	0.1766	0.1741	0.1291	0.1268	0.0095	0.0479	0.0990
$(1^L,1^F)_s,(2^L,1^F)_a$	0.2371	0.1766	0.1741	0.1291	0.0990	0.0479	0.0095	0.1268
$(1^L,1^F)_s,(2^L,2^F)_a$	0.2371	0.1766	0.1741	0.1291	0.0479	0.0990	0.0095	0.1268
$(1^L,2^F)_s,(1^L,1^F)_a$	0.2371	0.1766	0.1741	0.1291	0.0095	0.0990	0.0479	0.1268
$(1^L,2^F)_s,(1^L,2^F)_a$	0.2371	0.1766	0.1741	0.1268	0.0095	0.0990	0.0479	0.1291
$(1^L,2^F)_s,(2^L,1^F)_a$	0.2371	0.1766	0.1741	0.0990	0.0095	0.1268	0.0479	0.1291
$(1^L,2^F)_s,(2^L,2^F)_a$	0.2371	0.1766	0.1268	0.1291	0.0095	0.0990	0.0479	0.1741
$(2^L,1^F)_s,(1^L,1^F)_a$	0.2371	0.1766	0.1268	0.0990	0.0095	0.1291	0.0479	0.1741
$(2^L,1^F)_s,(1^L,2^F)_a$	0.2371	0.1766	0.0990	0.1268	0.0095	0.1291	0.0479	0.1741
$(2^L,1^F)_s,(2^L,1^F)_a$	0.2371	0.1766	0.0479	0.1291	0.0095	0.1268	0.0990	0.1741
$(2^L,1^F)_s,(2^L,2^F)_a$	0.2371	0.1766	0.0479	0.1268	0.0095	0.1291	0.0990	0.1741
$(2^L,2^F)_s,(1^L,1^F)_a$	0.2371	0.1766	0.0095	0.1268	0.0479	0.1291	0.0990	0.1741
$(2^L,2^F)_s,(1^L,2^F)_a$	0.2371	0.1741	0.0095	0.1291	0.0479	0.1268	0.0990	0.1766
$(2^L,2^F)_s,(2^L,1^F)_a$	0.2371	0.1741	0.0095	0.1268	0.0479	0.1291	0.0990	0.1766
$(2^L,2^F)_s,(2^L,2^F)_a$	0.2371	0.1291	0.0095	0.1268	0.0479	0.1741	0.0990	0.1766

,

$P(s(t+1), z^F(t+1)|s(t), \pi^L, \pi^F) =$

	$1^L_s,1^F_z,1^F_s$	$1^L_s,1^F_z,2^F_s$	$1^L_s,2^F_z,1^F_s$	$1^L_s,2^F_z,2^F_s$	$2^L_s,1^F_z,1^F_s$	$2^L_s,1^F_z,2^F_s$	$2^L_s,2^F_z,1^F_s$	$2^L_s,2^F_z,2^F_s$
$(1^L,1^F)_s,(1^L,1^F)_a$	0.21	0.17	0.18	0.12	0.13	0.07	0.05	0.07
$(1^L,1^F)_s,(1^L,2^F)_a$	0.21	0.17	0.18	0.12	0.13	0.02	0.04	0.13
$(1^L,1^F)_s,(2^L,1^F)_a$	0.21	0.17	0.18	0.12	0.11	0.05	0.01	0.15
$(1^L,1^F)_s,(2^L,2^F)_a$	0.21	0.17	0.18	0.12	0.06	0.10	0.01	0.15
$(1^L,2^F)_s,(1^L,1^F)_a$	0.21	0.17	0.18	0.12	0.02	0.10	0.05	0.15
$(1^L,2^F)_s,(1^L,2^F)_a$	0.21	0.17	0.18	0.11	0.02	0.10	0.04	0.17
$(1^L,2^F)_s,(2^L,1^F)_a$	0.21	0.17	0.18	0.10	0.01	0.11	0.04	0.18
$(1^L,2^F)_s,(2^L,2^F)_a$	0.21	0.17	0.14	0.12	0.01	0.10	0.05	0.20
$(2^L,1^F)_s,(1^L,1^F)_a$	0.21	0.17	0.14	0.09	0.01	0.13	0.05	0.20
$(2^L,1^F)_s,(1^L,2^F)_a$	0.21	0.17	0.11	0.12	0.01	0.13	0.05	0.20
$(2^L,1^F)_s,(2^L,1^F)_a$	0.21	0.17	0.06	0.13	0.01	0.12	0.10	0.20
$(2^L,1^F)_s,(2^L,2^F)_a$	0.21	0.17	0.06	0.12	0.01	0.13	0.10	0.20
$(2^L,2^F)_s,(1^L,1^F)_a$	0.21	0.17	0.02	0.12	0.05	0.13	0.10	0.20
$(2^L,2^F)_s,(1^L,2^F)_a$	0.21	0.16	0.02	0.12	0.045	0.13	0.10	0.215
$(2^L,2^F)_s,(2^L,1^F)_a$	0.21	0.16	0.02	0.11	0.04	0.135	0.10	0.225
$(2^L,2^F)_s,(2^L,2^F)_a$	0.21	0.13	0.01	0.13	0.03	0.165	0.10	0.225

$$Q^{L*} \in K(\pi^L, \pi^F) \cap K(\pi^L, \rho^F), Q^{L*} = \begin{bmatrix} 0.6 & 0.4 \\ 0.4 & 0.6 \end{bmatrix}$$

- reward structure:

  $R^F(d^F(t,\tau), \pi^L, \rho^F) = [2.0944, -10, 9.1798, -10, 9.1768, -10, 9.1858,$

  $-10, -10, 9.3521, -10, 9.3522, -10, 9.3620, -10, 2.8030]$

  $R^F(d^F(t,\tau), \pi^L, \pi^F) = [3.3540, 10, -9.3656, 10, -9.3656, 10, -9.3656, 10,$

  $10, -9.3656, 10, -9.3656, 10, -9.3656, 10, -9.3656]$

  $R^F(d^F(t,\tau), \rho^L, \rho^{F'}) = -\infty, \forall \rho^{F'} \in \Pi^F$

$R^L(d^F(t,\tau),\pi^L,\rho^F) = [2.9118, 2.8947, 2.8725, 2.8715, 2.7401, 2.7174, 2.4442, 2.4008,$
$1.8971, 1.6406, 1.4561, 0.8355, 0.4728, 0.4257, 0.3810, 0.2926]$
$R^L(d^F(t,\tau),\pi^L,\pi^F) = [2.0224, 1.9607, 1.9596, 1.9568, 1.9523, 1.9479, 1.7010, 1.6948,$
$1.2183, 0.9849, 0.8006, 0.4137, 0.3452, 0.2608, 0.2545, 0.0806]$

The parameters in Example 3.3 are:

(i) example(a):

- transition probabilities:

$$P^L(1) = \begin{bmatrix} 0.3202 & 0.6798 \\ 0.3044 & 0.6956 \end{bmatrix}, P^L(2) = \begin{bmatrix} 0.7624 & 0.2376 \\ 0.3790 & 0.6210 \end{bmatrix}$$

$$P^F(1) = \begin{bmatrix} 0.2593 & 0.7407 \\ 0.6356 & 0.3644 \end{bmatrix}, P^F(2) = \begin{bmatrix} 0.5221 & 0.4779 \\ 0.0994 & 0.9006 \end{bmatrix}$$

- reward structure $r^k(s^L, s^F, a^L, a^F), k \in \{L, F\}$:

$$r^F = r^L = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{c} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \end{array} \\ \begin{bmatrix} 9.3942 & 3.5084 & 6.4232 & 0.2711 \\ 4.1759 & 5.0135 & 9.2925 & 2.2993 \\ 2.9319 & 0.8495 & 8.3598 & 6.4256 \\ 0.0611 & 1.7647 & 3.3969 & 5.4417 \end{bmatrix} \end{array}$$

(ii) example(b):

- transition probabilities:

$$P^L(1) = \begin{bmatrix} 0.3277 & 0.6723 \\ 0.9623 & 0.0377 \end{bmatrix}, P^L(2) = \begin{bmatrix} 0.4723 & 0.5277 \\ 0.4469 & 0.5531 \end{bmatrix}$$

$$P^F(1) = \begin{bmatrix} 0.8815 & 0.1185 \\ 0.6147 & 0.3853 \end{bmatrix}, P^F(2) = \begin{bmatrix} 0.0641 & 0.9359 \\ 0.2062 & 0.7938 \end{bmatrix}$$

- reward structure $r^k(s^L, s^F, a^L, a^F), k \in \{L, F\}$:

$$r^F = r^L = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{c} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \end{array} \\ \begin{bmatrix} 6.0012 & 8.8066 & 5.3363 & 4.4058 \\ 9.8051 & 0.1065 & 0.6272 & 0.9891 \\ 7.6433 & 5.7154 & 5.6470 & 1.2712 \\ 7.1890 & 9.7160 & 2.2813 & 8.0163 \end{bmatrix} \end{array}$$

(iii) example(c):

- transition probabilities:

$$P^L(1) = \begin{bmatrix} 0.7657 & 0.2343 \\ 0.9270 & 0.0730 \end{bmatrix}, P^L(2) = \begin{bmatrix} 0.5570 & 0.4430 \\ 0.8113 & 0.1887 \end{bmatrix}$$

$$P^F(1) = \begin{bmatrix} 0.3594 & 0.6406 \\ 0.0007 & 0.9993 \end{bmatrix}, P^F(2) = \begin{bmatrix} 0.4647 & 0.5353 \\ 0.7964 & 0.2036 \end{bmatrix}$$

- reward structure $r^k(s^L, s^F, a^L, a^F), k \in \{L, F\}$:

$$r^F = r^L = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{c} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \end{array} \\ \begin{bmatrix} 3.3740 & 8.2159 & 1.6372 & 3.5419 \\ 7.5482 & 8.1603 & 5.8402 & 3.2979 \\ 7.0649 & 8.1617 & 9.0912 & 7.2706 \\ 4.4256 & 5.0106 & 2.8963 & 6.9836 \end{bmatrix} \end{array}$$

(iv) example(d):

- transition probabilities:

$$P^L(1) = \begin{bmatrix} 0.5983 & 0.4017 \\ 0.6592 & 0.3408 \end{bmatrix}, P^L(2) = \begin{bmatrix} 0.8785 & 0.1215 \\ 0.3068 & 0.6932 \end{bmatrix}$$

$$P^F(1) = \begin{bmatrix} 0.7036 & 0.2964 \\ 0.5885 & 0.4115 \end{bmatrix}, P^F(2) = \begin{bmatrix} 0.0593 & 0.9407 \\ 0.3593 & 0.6407 \end{bmatrix}$$

- reward structure $r^k(s^L, s^F, a^L, a^F), k \in \{L, F\}$:

$$r^F = r^L = \begin{array}{c} \\ (1^L,1^F)_s \\ (1^L,2^F)_s \\ (2^L,1^F)_s \\ (2^L,2^F)_s \end{array} \begin{array}{c} \begin{array}{cccc} (1^L,1^F)_a & (1^L,2^F)_a & (2^L,1^F)_a & (2^L,2^F)_a \end{array} \\ \begin{bmatrix} 7.3817 & 6.7738 & 4.4108 & 0.6868 \\ 9.0358 & 7.3192 & 3.5113 & 3.4176 \\ 7.2087 & 8.9864 & 9.3087 & 0.1376 \\ 6.4387 & 9.3377 & 9.5397 & 3.9611 \end{bmatrix} \end{array}$$